Evaluating Temporal Models at Matched Information

Anonymous Author
Anonymous Institution

Abstract

Temporal dependence makes raw sequence length a poor proxy for statistical information, yet temporal deep-learning benchmarks routinely compare regimes at fixed raw length N. We propose a dependence-aware evaluation protocol that matches an operational effective sample size N_{eff} across regimes: estimate the integrated autocorrelation time on a pilot sample, allocate raw lengths to a common information budget, verify the achieved N_{eff}, and report both the fixed-N and the matched view. Across more than 2,700 controlled runs spanning AR(1) and AR(2) processes, three architectures (TCN, LSTM, causal Transformer), and effective budgets from 50 to 4000, stronger dependence does not significantly increase train–test gaps in any matched cell; at scarce budgets the strongly dependent regime shows significantly smaller gaps, a contrast unchanged when every regime receives an identical number of optimizer updates, so it reflects information content rather than training exposure. A Beijing PM2.5 case study shows that which preprocessing appears best depends on the common information budget, an axis fixed-N evaluation cannot expose. We complement the protocol with a conservative generalization baseline for TCNs on exponentially β-mixing sequences, instantiating standard blocking with existing convolutional norm bounds; it controls population risk against an anchor empirical risk and serves as a worst-case reference for dependence and architecture scaling.

Preliminary work. Under review by AISTATS 2027. Do not distribute.

1 INTRODUCTION

Modern temporal architectures, including Temporal Convolutional Networks (TCNs) (Lea et al., 2017; Bai, 2018) and Transformer variants (Vaswani et al., 2017), are routinely evaluated on long dependent sequences (Lim et al., 2021; Oreshkin et al., 2019). Yet raw length N is not statistical information: strong dependence can make adjacent examples nearly redundant, and many routine benchmark decisions (sensor sampling rate, smoothing or downsampling, choice of dataset or operating regime) silently change temporal dependence. Two datasets with the same raw length can therefore carry very different amounts of statistical information, and a fixed-N comparison can attribute a difference in train–test gap to a modeling choice when it is partly an unequal information budget.

We focus on this evaluation problem. Effective sample size is classical in time-series statistics and Markov chain Monte Carlo (MCMC) diagnostics (Geyer, 1992; Sokal, 1997); our contribution is to make it an explicit matching criterion for temporal deep-learning benchmarks. We distinguish *standard evaluation*, which compares regimes at fixed raw length N, from *fair comparison*, which matches an operational effective sample size N_{eff} before comparison. The protocol is a four-step loop, summarized in the box below; a *pilot* is a preliminary sample of the same process used only for planning, in the sense of a pilot study, and the initial-positive-sequence (IPS) estimator is a classical truncation rule for the integrated autocorrelation time τ_{int} (Geyer, 1992).

The matched-N_{eff} protocol

1. **Estimate.** For each regime i, estimate the integrated autocorrelation time $\widehat{\tau}_i$ with the IPS estimator, on a pilot sample at the deployment sequence scale (or from a closed form when τ_{int} is known analytically).
2. **Allocate.** Choose raw lengths $N_i = \lceil N_\star \widehat{\tau}_i \rceil$ so that every regime targets the same effective budget $N_{\text{eff}} = N_\star$.
3. **Verify.** Re-estimate the achieved $\widehat{N}_{\text{eff}}$ on the realized experimental data; re-allocate once if it misses the target.
4. **Report.** Report results under both the fixed-N and the matched-N_{eff} view, with raw lengths, pilot estimates, and achieved budgets.

This helps benchmark designers tell whether a changed gap or ranking reflects the model or preprocessing choice or simply an unequal information budget. It is not blocked cross-validation, which asks how to score one dependent stream without leakage; matching asks whether different regimes receive the same budget. In controlled autoregressive sweeps, strongly dependent regimes need not show larger generalization gaps once the budget is matched; at scarce budgets their gaps are significantly smaller.

We complement the protocol with a conservative theoretical baseline, deliberately introducing no new theorem: the theory instantiates blocking/coupling reductions for mixing processes (Yu, 1994; Mohri and Rostamizadeh, 2008) with norm-based i.i.d. complexity bounds for convolutional networks that account for weight sharing (Golowich et al., 2018; Ledent et al., 2021), as a worst-case reference that makes the dependence and architecture scaling explicit for TCNs. The guarantee controls population risk against an *anchor* empirical risk (the empirical risk on a spaced subsample of time points), whose anchor count B is the theory-side effective sample size induced by blocking, distinct from the autocorrelation-based N_{eff} used operationally; we keep both explicit.

We make three contributions:

1. **Fair-comparison methodology for dependent sequences**: operational N_{eff} matching for comparing dependence regimes, sampling rates, or preprocessing choices that change temporal dependence, with a reporting checklist and a pilot-stage diagnostic for its mixing assumption.
2. **Empirical findings enabled by fair comparison.** In more than 2,700 controlled runs, fixed-N and matched-N_{eff} evaluation lead to different qualitative conclusions about dependence; the effect concentrates at scarce information budgets, survives an exposure control that fixes the number of optimizer updates, and holds for TCN, LSTM, and causal Transformer predictors. A Beijing PM2.5 diagnostic shows that which preprocessing appears best depends on the common budget level.
3. **A dependence-aware generalization baseline for TCNs**: an instantiation of prior blocking and convolutional complexity results for norm-controlled TCNs on exponentially β-mixing sequences, with explicit dependence on depth, kernel size, norm budget, and anchor count.

Section 2 reviews related work, Section 3 fixes notation and separates the anchor count B from operational N_{eff}, Section 4 presents the bound, Section 5 the empirical evidence, and the final section limitations and future work.

2 RELATED WORK

Temporal models and evaluation methodology. TCNs (Lea et al., 2017; Bai, 2018) and Transformer-style sequence models (Vaswani et al., 2017; Lim et al., 2021; Oreshkin et al., 2019) are routinely evaluated on long dependent sequences. Blocked/rolling validation and prequential evaluation estimate risk for a fixed temporal process (Bergmeir et al., 2018; Dawid, 1984), asking how to split or score one dependent stream without leakage; matched-N_{eff} evaluation instead asks whether runs with different dependence, preprocessing, or sampling receive the same statistical information budget, using classical effective sample size diagnostics (Geyer, 1992; Sokal, 1997) as matching variables rather than proposing a new dependence estimator.

Learning under dependence. Blocking/coupling reductions yield concentration and uniform convergence for stationary mixing processes, including empirical-process rates for β-mixing sequences (Yu, 1994), nonparametric prediction under dependence (Meir, 2000), and Rademacher-complexity bounds for mixing samples (Mohri and Rostamizadeh, 2008); other approaches use stability (Mohri and Rostamizadeh, 2009), discrepancy (Kuznetsov and Mohri, 2014), PAC-Bayes (Alquier and Guedj, 2018), or online-to-PAC reductions (Abeles et al., 2024). We focus on β-mixing because absolute regularity gives the total-variation coupling used for anchors.

Norm-based bounds for neural networks and convolutions. Existing sequence-model bounds study RNN-style predictors under dependence (Kuznetsov and Mohri, 2014) or norm/stability perspectives for attention (Trauger and Tewari, 2024; Zhang et al.,

2024); our theoretical instantiation targets temporal convolutions. Parameter-count and VC-style bounds can be loose for modern neural networks (Bartlett et al., 1998, 2019). Norm-based bounds scale with products of layer norms and often improve depth dependence (Neyshabur et al., 2015; Bartlett et al., 2017; Golowich et al., 2018). For convolutional networks, the analysis must also account for weight sharing and structured linear operators: Long and Sedghi (2019) and Ledent et al. (2021) provide CNN bounds that avoid expanding convolutions into dense matrices, with later refinements studying filter-level norms and excess capacity (Galanti et al., 2023; Graf et al., 2022). We use this convolutional i.i.d. ingredient and lift it to dependent data via blocking.

3 PRELIMINARIES

We analyze temporal models trained on dependent data using three ingredients: (i) a dependence model (β-mixing), (ii) an i.i.d. capacity measure (Rademacher complexity), and (iii) an empirical information proxy for fair comparison (effective sample size N_{eff}).

Learning from a single dependent time series. Let $\{z_t\}_{t=1}^{N}$ be a finite realization of a stationary process with $z_t \in \mathbb{R}^n$. We consider one-step-ahead prediction with context length $q \geq 1$: $x_t = (z_{t-q+1}, \ldots, z_t) \in \mathbb{R}^{q\times n}$ and $y_t = z_{t+1} \in \mathbb{R}^n$ for $t = q, \ldots, N-1$, yielding $m = N - q$ dependent examples.

Given a predictor $f : \mathbb{R}^{q\times n} \to \mathbb{R}^n$ and loss $\ell : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, the population risk and empirical risk are $\mathcal{L}(f) = \mathbb{E}\big[\ell(f(x_t), y_t)\big]$, $\widehat{\mathcal{L}}_m(f) = \frac{1}{m}\sum_{t=q}^{N-1} \ell\big(f(x_t), y_t\big)$, where the expectation is w.r.t. the stationary law of the process. Practitioners usually report gaps against $\widehat{\mathcal{L}}_m$; our formal bound controls an anchor empirical risk constructed by blocking; Section 4 spells out the distinction.

Stationary β-mixing processes. Stationarity keeps finite-window distributions time-invariant. We quantify dependence by β-mixing. Let $U_t = (x_t, y_t)$ denote the example process, $\mathcal{F}_a^b = \sigma(U_s : a \leq s \leq b)$, $\mathcal{F}_{-\infty}^t = \sigma(U_s : s \leq t)$, and $\mathcal{F}_{t+k}^{\infty} = \sigma(U_s : s \geq t+k)$. Throughout, $\beta(\cdot)$ refers to the β-mixing coefficients of $\{U_t\}$: $\beta(k) = \sup_{t\geq 1} \mathbb{E}\Big[\sup_{A\in\mathcal{F}_{t+k}^{\infty}} \big|\mathbb{P}(A \mid \mathcal{F}_{-\infty}^t) - \mathbb{P}(A)\big|\Big]$. A small $\beta(k)$ means observations separated by k steps are nearly independent.

Assumption 3.1 (Exponential β-mixing)**.** There exist constants $C_0, c_0 > 0$ such that for all $k \geq 0$, $\beta(k) \leq C_0 e^{-c_0 k}$.

This condition justifies the blocking/coupling reduction that is the core technical step of Section 4.

Remark 3.2 (Mixing of windowed examples). Let $\beta_z(\cdot)$ denote the β-mixing coefficients of the raw process $\{z_t\}$. Since each example $U_t = (x_t, y_t)$ depends only on $(z_{t-q+1}, \ldots, z_{t+1})$, the example process $\{U_t\}$ is also β-mixing and satisfies, for all $k > q$, $\beta_U(k) \leq \beta_z(k - q)$. Throughout this paper, $\beta(\cdot)$ refers to the mixing coefficients of the windowed example process $\{U_t\}$, and the delay parameter d counts steps in the example index (i.e., anchors U_i and U_j are separated by $|i - j|$ steps).

Starting from raw-process coefficients thus incurs a fixed q-offset in the delay choice, changing constants in the anchor count but not the $B = \Theta(m/\log m)$ scaling; Appendix B.6 gives the calibration for a first-order autoregressive (AR(1)) process, writing AR(k) for an autoregressive process of order k.

Rademacher complexity (i.i.d. ingredient). For an i.i.d. sample $S = \{u_i\}_{i=1}^{M}$ and vector-valued predictors we use the standard analogue $\widehat{\mathfrak{R}}_S^{\text{vec}}(\mathcal{F}) = \frac{1}{M}\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \sum_i \langle \sigma_i, f(u_i)\rangle$, where $\sigma_i \in \{\pm 1\}^n$ have i.i.d. Rademacher coordinates; vector contraction (Maurer, 2016; Ledoux and Talagrand, 2013) gives $\mathfrak{R}_M(\ell \circ \mathcal{F}) \leq C_{\text{vc}} L\, \mathfrak{R}_M^{\text{vec}}(\mathcal{F})$ for an L-Lipschitz loss. We apply this after constructing approximately independent anchors.

Effective sample size and fair comparison. For dependent data, the raw number of examples m can misstate information content. We use N_{eff} as an *operational matching variable*: two regimes are compared at the same estimated information budget rather than the same raw length. We do not propose a new estimator; we use classical autocorrelation-function (ACF) diagnostics and the initial-positive-sequence (IPS) truncation rule to make the protocol explicit. When τ_{int} is not known analytically, it is estimated on a *pilot* sample of the same process, used only for allocation; the achieved N_{eff} is then verified on the realized experimental data (the allocate-verify-adjust loop of the protocol box in the introduction).

For model comparisons on one fixed dataset and split, N_{eff} is shared by construction and should simply be reported once; the protocol matters when the data regime changes (dependence strength, sampling rate, preprocessing, dataset).

Relationship between theory and empirical calibration. The theory in Section 4 is controlled by the *anchor count* $B = \lfloor m/(d+1)\rfloor$ induced by blocking, which scales as $\Theta(m/\log m)$ under our delay choice; the experiments match the ACF/IPS N_{eff} (Section 5), which captures autocorrelation-induced variance inflation. The two are related but distinct, and we keep both explicit (Appendix B.6).

Model class: causal TCNs and norm control.

We focus on causal TCN predictors built from 1D convolutions and ReLU activations ($\sigma(x) = \max(0, x)$ elementwise; 1-Lipschitz with $\sigma(0) = 0$, properties used in Section 4). Let a depth-D TCN have convolutional weight tensors $W^{(\ell)} \in \mathbb{R}^{C_\ell \times C_{\ell-1} \times p}$, $\ell = 1, \dots, D$, with kernel size p and channels C_ℓ. To control capacity, we use a filter-group norm (an $\ell_{2,1}$-type norm over output filters): $\|W^{(\ell)}\|_{2,1} = \sum_{j=1}^{C_\ell} \|W^{(\ell)}_{j,:,:}\|_2$.

We impose $\|W^{(\ell)}\|_{2,1} \le M^{(\ell)}$ and write $R = \prod_{\ell=1}^{D} M^{(\ell)}$ for the norm-budget product appearing in architecture-aware complexity bounds. **Architecture conventions:** The input $x_t \in \mathbb{R}^{q \times n}$ is a sequence of q time steps with n features, so the layer-0 channel dimension is $C_0 = n$. A 1D causal convolution with kernel size p uses positions $i - p + 1, \dots, i$ to compute output position i. The hypothesis class is

$$\begin{aligned}\mathcal{F}_{D,p,R,B_f} = \{f_W : f_W \text{ a depth-}D \text{ causal TCN},\\ \text{kernel size } p,\ \|W^{(\ell)}\|_{2,1} \le M^{(\ell)}\ \forall \ell,\\ \textstyle\prod_\ell M^{(\ell)} \le R,\\ \|f_W(x)\|_2 \le B_f\ \forall \|x\|_F \le B_x\}.\end{aligned} \tag{1}$$

The output bound B_f is enforced via an output clipping layer $\mathrm{clip}_{B_f}(y) = y \cdot \min(1, B_f/\|y\|_2)$, with $\mathrm{clip}_{B_f}(0) = 0$. Clipping is 1-Lipschitz; the contraction constant it introduces is absorbed into C'. We write $\mathcal{F}_{D,p,R}$ when B_f is clear.

Boundedness assumptions

Assumption 3.3 (Bounded inputs, outputs, and targets)**.** There exist constants $B_x, B_y, B_f > 0$ such that: (i) $\|x_t\|_F \le B_x$ almost surely (equivalently $\|\mathrm{vec}(x_t)\|_2 \le B_x$), (ii) $\|y_t\|_2 \le B_y$ almost surely. The output bound (iii) $\|f(x)\|_2 \le B_f$ for all $\|x\|_F \le B_x$ is enforced by the hypothesis class definition equation 1 via output clipping, not assumed on trained models.

Remark 3.4 (Dependence of B_x on context window length). If $\|z_t\|_2 \le b_z$ for all t, then $\|x_t\|_F \le \sqrt{q}\, b_z$: weight sharing prevents explicit scaling with raw sequence length, but B_x depends on q. In our experiments q is constant across dependence regimes within every comparison ($q = 32$ in the main sweep; $q \in \{8, 16\}$ in the small-budget sweep, where shorter raw sequences require shorter windows), so this dependence is fixed in the reported comparisons.

Assumption 3.5 (Lipschitz loss)**.** For each fixed y, the map $\hat{y} \mapsto \ell(\hat{y}, y)$ is L-Lipschitz in $\|\cdot\|_2$.

Under Assumption 3.3, squared loss is bounded by $(B_f + B_y)^2$ and L-Lipschitz with $L = 2(B_f + B_y)$; results stated for losses in $[0, 1]$ are applied to the normalized loss $\bar{\ell} = \ell/(B_f + B_y)^2$ and rescaled.

Notation Table 2 summarizes the main notation.

4 CONSERVATIVE TCN BOUND UNDER β-MIXING

This section gives a conservative generalization baseline for temporal convolutional predictors trained on one dependent sequence. We do not introduce a new blocking theorem or CNN complexity bound: the dependent-data step follows Yu/Mohri–Rostamizadeh blocking, and the i.i.d. architectural ingredient comes from norm-based convolutional-network bounds; the contribution is the TCN instantiation and the explicit separation between theory-side anchor risk and the full empirical risk used in experiments. We work under Assumption 3.1 and the boundedness/Lipschitz conditions from Section 3, with ℓ bounded in $[0, 1]$ and L-Lipschitz (normalized squared loss as in Section 3).

Blocking and coupling. Recall the example process $U_t = (x_t, y_t)$ from Section 3. For notation matching the appendix proofs, write $Z_t = U_t$ and index the m windowed examples as $(Z_t)_{t=1}^m$. Let $d \ge 1$ be a spacing parameter. Partition the first $B(d+1)$ indices into consecutive blocks of length $d+1$: $I_j = \{(j-1)(d+1)+1, \dots, j(d+1)\}$, $j = 1, \dots, B$, where $B = \lfloor m/(d+1) \rfloor$. From each block take $A_j = Z_{(j-1)(d+1)+1}$. Anchors are separated by $d+1$ time steps, so their dependence is controlled by $\beta(d+1)$. Appendix Figure 16 gives the schematic; the main trade-off is that larger d reduces residual dependence but also reduces B (Remark 3.2 handles raw-process coefficients).

Lemma 4.1 (Coupling of block anchors)**.** *Under Assumption 3.1, the joint law of anchors is close (in total variation) to a product law:* $\left\| P_{(A_1,\dots,A_B)} - \bigotimes_{j=1}^{B} P_{A_j} \right\|_{\mathrm{TV}} \le (B-1)\,\beta(d+1) \le B\,\beta(d+1)$.

The proof (Appendix D.2) telescopes over the $B-1$ block interfaces, each bounded by $\beta(d+1)$.

From dependent anchors to an i.i.d. generalization bound. Define the *anchor empirical risk* $\widehat{\mathcal{L}}^{\mathrm{anc}}_B(f) = \frac{1}{B}\sum_{j=1}^{B} \ell\big(f(x(A_j)), y(A_j)\big)$, where $x(A_j), y(A_j)$ are the input/label components of A_j. By stationarity, $\mathbb{E}[\widehat{\mathcal{L}}^{\mathrm{anc}}_B(f)] = \mathcal{L}(f)$ for every fixed f. Lemma 4.1 couples $(A_1, \dots, A_B)$ to an i.i.d. sample with matching marginals up to total variation $(B-1)\beta(d+1)$, so i.i.d. generalization bounds transfer to anchors with this extra failure probability.

Theorem 4.2 (Generic β-mixing generalization via blocking)**.** *Let $(Z_t)_{t=1}^m$ be strictly stationary and β-mixing with coefficients $\beta(\cdot)$, where each $Z_t = (x_t, y_t) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\ell : \mathbb{R}^n \times \mathbb{R}^n \to [0, 1]$ be a loss, and define the loss-composed class $\ell \circ \mathcal{F} = \{z = (x, y) \mapsto \ell(f(x), y) : f \in \mathcal{F}\}$. Fix $d \ge 1$ and define $B = \lfloor m/(d+1) \rfloor$ anchors $A_j = Z_{1+(j-1)(d+1)}$, $j = 1, \dots, B$. Then*

for any $\delta \in (0,1)$, with probability at least $1-\delta-(B-1)\beta(d+1)$,

$$\sup_{f\in\mathcal{F}}\left|\mathcal{L}(f)-\widehat{\mathcal{L}}_B^{\mathrm{anc}}(f)\right| \;\le\; 2\,\mathfrak{R}_B(\ell\circ\mathcal{F}) \;+\; 3\sqrt{\frac{\log(2/\delta)}{2B}}.$$

Moreover, if $\ell(\cdot,y)$ is L-Lipschitz (in $\|\cdot\|_2$) for all y, then for a universal vector-contraction constant C_{vc},

$$\mathfrak{R}_B(\ell\circ\mathcal{F}) \;\le\; C_{\mathrm{vc}} L\,\mathfrak{R}_B^{\mathrm{vec}}(\mathcal{F}).$$

The proof (Appendix D.3) couples the anchors to an i.i.d. sample with matching marginals, paying $(B-1)\beta(d+1)$ in total variation, and applies standard symmetrization.

Choosing the spacing d. Under exponential mixing of the windowed example process (Assumption 3.1), setting $d=\lceil \log m/c_0\rceil$ gives $\beta(d+1)\le C_0/m$, coupling slack $B\beta(d+1)=O(1/\log m)$, and a dominant concentration rate of $O\big(\sqrt{\log m/m}\big)$; for a prescribed slack η, the closed-form choice $d\ge c_0^{-1}\log(C_0 m/\eta)$ suffices since $B\le m$.

Instantiating with causal TCNs under filter-group norm control. We now instantiate Theorem 4.2 with causal TCNs. Let each convolutional layer satisfy the filter-group ($\ell_{2,1}$) constraint $\|W^{(\ell)}\|_{2,1}\le M^{(\ell)}$ and denote the product budget $R=\prod_{\ell=1}^{D} M^{(\ell)}$.

Lemma 4.3 (Imported CNN complexity ingredient for norm-controlled TCNs)**.** *Let $\mathcal{F}_{D,p,R}$ be depth-D causal TCNs with kernel size p, ReLU activations, under $\ell_{2,1}$ filter-group norm control with product budget $R=\prod_{\ell=1}^{D} M^{(\ell)}$. Assume inputs satisfy $\|x\|_F\le B_x$ almost surely. Then for an i.i.d. sample of size B there is an architecture/logarithmic factor $\Lambda_{\mathrm{TCN}}(B,\mathbf{C}_{\mathrm{ch}},n,q,p)$, inherited from the convolutional covering bounds of Ledent et al. (2021), such that $\mathfrak{R}_B^{\mathrm{vec}}(\mathcal{F}_{D,p,R}) \;\le\; \Lambda_{\mathrm{TCN}}(B,\mathbf{C}_{\mathrm{ch}},n,q,p)\,\frac{R\,B_x\sqrt{D\log(2p)}}{\sqrt{B}}$, where $\mathbf{C}_{\mathrm{ch}}=(C_0,\dots,C_D)$ denotes channel widths. For fixed architecture, Λ_{TCN} is polylogarithmic in the sample/patch counts and does not scale linearly with the raw sequence length. The dependence on the context length enters through B_x and fixed-architecture factors (Remark 3.4).*

The proof (Appendix D.5) specializes the convolutional covering/Rademacher machinery of Ledent et al. (2021), whose covers count each filter once rather than expanding layers into dense Toeplitz matrices, to fixed-window 1D causal convolutions, and verifies the patch-wise Cauchy–Schwarz control needed for our $\ell_{2,1}$ filter convention; we do not claim global spectral domination or a no-overlap Frobenius Lipschitz bound.

Main bound. Combining Theorem 4.2 with Lemma 4.3 yields the following baseline.

Theorem 4.4 (Architecture-aware baseline under exponential β-mixing)**.** *Let Assumptions 3.1, 3.3, and 3.5 hold. Let $\mathcal{F}_{D,p,R}$ (shorthand for $\mathcal{F}_{D,p,R,B_f}$) be the norm-controlled TCN class in Lemma 4.3, and let $B=\lfloor m/(d+1)\rfloor$ be the number of anchors. Then with probability at least $1-\delta-(B-1)\beta(d+1)$,*

$$\sup_{f\in\mathcal{F}_{D,p,R}}\left|\mathcal{L}(f)-\widehat{\mathcal{L}}_B^{\mathrm{anc}}(f)\right| \le C' L\Lambda_{\mathrm{TCN}} R B_x\frac{\sqrt{D\log(2p)}}{\sqrt{B}} + 3\sqrt{\frac{\log(2/\delta)}{2B}}. \tag{2}$$

where $\Lambda_{\mathrm{TCN}}=\Lambda_{\mathrm{TCN}}(B,\mathbf{C}_{\mathrm{ch}},n,q,p)$ is the architecture/log factor from Lemma 4.3, and C' is a universal contraction/concentration constant independent of the mixing parameters C_0,c_0 (though the delay choice $d^*=\lceil\log m/c_0\rceil$ does depend on c_0).

Corollary 4.5 (Bounded squared loss)**.** *Under Assumption 3.3, let $\ell_{\mathrm{sq}}(\hat y,y)=\|\hat y-y\|_2^2$ and $A=(B_f+B_y)^2$. Applying Theorem 4.4 to $\bar\ell=\ell_{\mathrm{sq}}/A$ and rescaling gives the same probability statement with*

$$\begin{aligned}&\sup_{f\in\mathcal{F}_{D,p,R}}\left|\mathcal{L}_{\mathrm{sq}}(f)-\widehat{\mathcal{L}}_{B,\mathrm{sq}}^{\mathrm{anc}}(f)\right|\\ &\le 2C'(B_f+B_y)\Lambda_{\mathrm{TCN}} R B_x\frac{\sqrt{D\log(2p)}}{\sqrt{B}}\\ &\quad+3A\sqrt{\frac{\log(2/\delta)}{2B}}.\end{aligned}$$

Remark 4.6 (Anchor vs. full empirical risk). Theorems 4.2 and 4.4 control the gap between the population risk and the *anchor* empirical risk $\widehat{\mathcal{L}}_B^{\mathrm{anc}}(f)$, not the full empirical risk $\widehat{\mathcal{L}}_m(f)$. Both estimators are unbiased under stationarity, but without additional structure their difference admits only the deterministic bound $1-B/m$ (Lemma D.4, Appendix D), which is not small when $d=\Theta(\log m)$; accordingly, the theory is stated for anchor risk. Lemma D.8 gives a refinement under within-block smoothness, and Appendix D.4 a block-averaged estimator compatible with blocking.

Remark 4.7 (Nature of the bound; anchors in practice). Theorem 4.4 is a *uniform convergence* guarantee: it holds simultaneously for all $f\in\mathcal{F}_{D,p,R}$, regardless of how f is selected, so algorithm-specific stability or implicit-regularization analyses could be tighter. Anchors are a proof device: our experiments train on all available data and report full train–test gaps, which are descriptive diagnostics, not formal consequences of the theorem.

The proof of Theorem 4.4 (Appendix D.7) combines Theorem 4.2, vector contraction, and Lemma 4.3, choosing $d=\lceil\log m/c_0\rceil$ so that the slack is $O(1/\log m)$ while $B=\Theta(m/\log m)$.

Theorem 4.4 is conservative: it establishes learnability and makes explicit how dependence (through $\log m$) and architecture (through D, p, R) enter. It complements the fair-comparison methodology by clarifying worst-case scaling, but it is not intended to predict empirical rates. Appendix B.12 gives an architectural-factor sanity check against a generic dependence-aware plug-in bound.

5 EXPERIMENTS

We evaluate the matched-N_{eff} protocol and use the TCN bound as a conservative reference. **The main contribution here is methodological:** when information content is held fixed, the apparent effect of dependence on generalization can change relative to fixed-length evaluation. We therefore foreground the operational protocol and treat the bound as a calibration lens. The evidence spans more than 2,700 training runs (matched AR(1) sweeps at rich and scarce budgets, exposure controls, two further architectures, and the Beijing panel).

Evaluation metric and experimental grid. Across all experiments we report the empirical generalization gap $\text{Gap}(f) = \widehat{\mathcal{L}}_{\text{test}}(f) - \widehat{\mathcal{L}}_{\text{train}}(f)$, where $\widehat{\mathcal{L}}$ is mean squared error (MSE). These gaps use the full train/test splits; the theorem controls anchor empirical risk, so bound curves below are diagnostics rather than formal guarantees for the plotted train–test gap. Slightly negative signed gaps are interpreted as essentially zero; we also report $|\text{Gap}(f)|$ in multi-slice diagnostics. Our headline statistic is the *strong-minus-weak contrast*: $|\text{Gap}|$ at $\rho = 0.8$ minus $|\text{Gap}|$ at $\rho = 0.2$ within a matched slice, paired by depth and trial, with paired-bootstrap 95% confidence intervals; negative values mean the strongly dependent regime has *smaller* gaps at matched information.

Train/validation/test splitting. Because examples are temporally dependent, we split each sequence *chronologically* into train, validation, and test segments (no shuffling). Windowed examples (x_t, y_t) are constructed *within* each split so that a validation or test window never shares raw time points with earlier splits; remaining boundary dependence is intrinsic to chronological forecasting. Each series is standardized with train-segment mean and variance before windowing, so MSE gaps are on comparable scales across ρ despite the AR(1) target variance growing as $1/(1-\rho^2)$. Early stopping and learning-rate scheduling use validation loss, while the test segment is used only for final reporting.

Experimental setup. For synthetic AR(1), we use $n = 1$, context $q = 32$, channels $C = 32$, kernel size $p = 3$, ReLU activations, and hidden-layer BatchNorm; training uses Adam with weight decay, gradient clipping, a filter-norm projection, and validation early stopping (patience 20 within a fixed 100-epoch budget, best-validation weights restored). Two variants below remove stopping-rule effects: a *no-early-stopping* variant disables the stopping rule so every run trains for exactly the full 100-epoch budget, and a *fixed-update* variant performs exactly the same number of Adam updates in every regime, with no validation-based stopping and no checkpoint selection, evaluating the final weights. The main sweep covers $N_{\text{eff}} \in \{500, 1000, 2000, 4000\}$, $\rho \in \{0.2, 0.4, 0.6, 0.8\}$, depths $D \in \{2, 4, 6, 8\}$, and 10 trials per depth (640 runs); the scarce-budget sweep covers $N_{\text{eff}} \in \{50, 100, 200\}$ with 20 trials and depths $\{2, 4\}$ (480 runs per stopping variant, 960 total), shortening the context window to $q = 16$ (targets 100, 200) and $q = 8$ (target 50), constant across ρ within each target, because the weak-dependence arm at $N_{\text{eff}} = 50$ has only 75 raw points. Appendices B.1–B.2 and C give all remaining hyperparameters and full summaries.

The formal guarantee applies to bounded, clipped, norm-controlled causal ReLU TCNs without BatchNorm; because the AR responses are Gaussian, no synthetic experiment is claimed to instantiate the theorem exactly, and the sweep uses the bound only as a post-hoc conservative diagnostic (a no-BatchNorm/clipped ablation appears in Appendix B.5). The diagnostic remains far above measured gaps: regressing signed gaps on the norm proxy times $\sqrt{D \log(2p) \log N / N}$ gives $R^2 \approx 0.01$ (Appendix B.13), consistent with a worst-case reference rather than a calibrated predictor.

For real-signal diagnostics, we use the Beijing Multi-Site Air Quality dataset (Zhang et al., 2017) and the MIT-BIH Arrhythmia Database (Goldberger et al., 2000). Beijing PM2.5 is used below as a matched-IPS protocol demonstration; fixed-raw-N PhysioNet scaling plots are deferred to Appendix B.8.

Fair comparison protocol. Stronger dependence reduces the number of effectively independent observations at fixed raw N; to separate "structure" from "information," we match **effective sample size**.

For an AR(1) process with lag-1 correlation ρ, we use the standard approximation (Wilks, 2011)

$$N_{\text{eff}} \approx N \cdot \frac{1-\rho}{1+\rho}, \quad \text{hence} \quad N(\rho) = \left\lfloor N_{\text{eff}} \cdot \frac{1+\rho}{1-\rho} \right\rfloor.$$

Thus $\rho = 0.2$ and $\rho = 0.8$ require about 3,000 vs. 18,000 raw points at $N_{\text{eff}} = 2000$; conversely, at fixed raw $N = 4096$, $\rho = 0.2$ carries $N_{\text{eff}} \approx 2731$ while $\rho = 0.8$ carries only ≈ 455, a roughly $6\times$ information-budget imbalance that matching removes. Matching controls a first-order information budget, not raw exposure, optimizer steps, or task smoothness; train–test gaps and

the proof-side anchor count B are measured separately. For non-oracle settings, we estimate $\widehat{\tau}_{\text{int}}$ from pilot sequences, set $N_i = \lceil N_\star \widehat{\tau}_i \rceil$, and verify achieved $\widehat{N}_{\text{eff}}$, following the protocol box in the introduction. The empirical question is whether learned gaps retain a systematic dependence on ρ after matching; a falsifying outcome would be a monotone increase of |Gap| with ρ across matched slices.

Fair comparison results: separating information from structure.

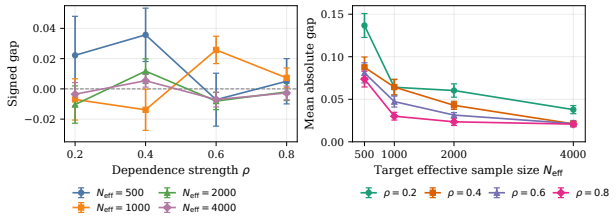


Figure 1: **Fair comparison at matched effective sample size** (640-run AR(1) sweep). Signed gaps remain centered near zero; absolute gaps decrease with target N_{eff} and do not increase with dependence at matched information. Points aggregate 40 trial-depth cells.

Figure 1 reports controlled comparisons where each vertical slice fixes the **same effective information content** and varies temporal dependence; the relative ordering of gaps across ρ would be mischaracterized under fixed-N evaluation. Appendix B.11 reports depth, anchor/full-risk, and bound diagnostics; Appendices B.1, B.6, and B.13 give the calibration details.

Key empirical diagnostic under fair comparison. At fixed $N_{\text{eff}} = 2000$, mean signed gaps stay near zero, while absolute gaps are 0.060 ± 0.008, 0.043 ± 0.005, 0.031 ± 0.003, and 0.024 ± 0.004 for $\rho = 0.2, 0.4, 0.6, 0.8$; the paired strong-minus-weak contrast is -0.037 ± 0.009. Across the full 640-run sweep, a paired bootstrap over the 160 target-depth-trial cells gives slope -0.061 for |Gap| versus ρ (95% CI $[-0.079, -0.044]$), while the signed-gap slope is -0.002 (95% CI $[-0.033, 0.029]$). Thus absolute gaps do not increase with ρ at matched information in this controlled grid, while signed gaps remain centered near zero. We treat this as a methodological diagnostic rather than a universal monotone law.

Scarce information budgets. The scarce-budget sweep ($N_{\text{eff}} \in \{50, 100, 200\}$) tests whether the diagnostic survives in the strong-overfitting regime that small budgets induce. The regime is genuinely unsaturated: mean |Gap| reaches 0.36, and at $N_{\text{eff}} = 100$, $\rho = 0.4$ the mean *signed* gap is $+0.21$ with 65% of runs positive, i.e., systematic overfitting. Table 1 reports the contrasts: exactly where the effect should be most visible if real, it is largest (|Gap| decreases monotonically in ρ at $N_{\text{eff}} = 100$, 0.354/0.343/0.256/0.152, contrast -0.203); at $N_{\text{eff}} = 50$ the contrast attenuates while remaining non-positive, all regimes being window-limited there. In no cell of either sweep does stronger dependence produce significantly larger gaps at matched information.

Table 1: Strong-minus-weak |Gap| contrasts at scarce matched budgets (20 seeds × depths $\{2, 4\}$ per cell; ES = early stopping). The ordering is unchanged without early stopping and under a fixed budget of 200 Adam updates.

N_{eff}	Early stop	No ES	Fixed upd.	95% CI (ES)
50	-0.036	-0.052	–	incl. 0
100	$\mathbf{-0.203}$	-0.210	$\mathbf{-0.201}$	$[-0.321, -0.085]$
200	-0.096	-0.096	–	$[-0.166, -0.026]$

Exposure control: fixed update budgets. Matched regimes have different raw lengths, so even a fixed epoch budget implies different numbers of gradient steps. The fixed-update variant removes this confounder: every run performs an identical number of Adam updates regardless of ρ. At $N_{\text{eff}} = 100$ with exactly 200 updates per run, the contrast is -0.201 ± 0.048 (CI $[-0.294, -0.107]$), unchanged from the standard recipe (-0.203): the scarce-budget effect reflects information content, not training exposure. At the rich budget $N_{\text{eff}} = 2000$, fixing 2,000 updates gives a null contrast of $+0.013$ (CI $[-0.035, +0.062]$), with an identical point estimate at 4,000 updates (CI $[-0.022, +0.048]$). The no-increase conclusion thus holds under fixed exposure at every budget: the negative contrast is information-driven and exposure-invariant where budgets are scarce, while at rich budgets the standard recipe's residual negative contrast is partly a training-duration effect, which is why the protocol prescribes reporting update counts whenever raw budgets differ (Appendix C.2).

Architecture generality. The mechanism is a property of the information budget, not of the predictor. Rerunning the matched protocol with a 2-layer LSTM and a 2-layer causal Transformer (10 seeds per cell, same recipe, targets $N_{\text{eff}} \in \{500, 2000\}$) yields negative contrasts in all four architecture-target cells, significant in three: Transformer -0.072 (CI $[-0.119, -0.024]$) and -0.038 (CI $[-0.056, -0.019]$); LSTM -0.036 (CI $[-0.070, -0.002]$) at target 2000 and -0.034 (CI includes 0) at target 500. In no cell do gaps increase with dependence at matched information, matching the

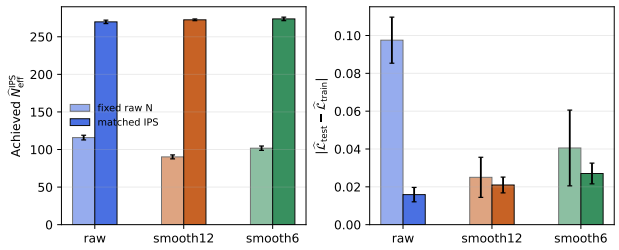


Figure 2: **Beijing matched-IPS diagnostic.** Fixed $N = 4096$ gives $\widehat{N}_{\text{eff}}^{\text{IPS}} \approx 116, 90, 102$ for raw, smooth12, and smooth6 (the panel axis order); IPS matching gives $\approx 270, 273, 274$. Gaps are descriptive because smoothing changes the task.

TCN result (Appendix C.3).

Estimating N_{eff} without oracle ρ, and a non-oracle AR(2) check. Beyond the oracle formula, a practical protocol estimates information content with the IPS estimator (Geyer, 1992; Sokal, 1997): truncate the paired autocorrelation sums at the first non-positive pair and set $\widehat{N}_{\text{eff}}^{(\text{IPS})} = N/\widehat{\tau}_{\text{int}}$. IPS recovers the known AR(1) autocorrelation time accurately (Appendix Table 4), averaging 1922/1958/2029/2029 across ρ at target 2000. As an end-to-end check, Appendix B.3 applies the full protocol to three stationary positive-dependence AR(2) processes: at fixed raw $N = 4096$, weak/medium/strong variants achieve $\widehat{N}_{\text{eff}}^{\text{IPS}} \approx 2110/796/310$, a large hidden imbalance; after pilot-IPS matching at target 2000, achieved sizes are $\approx 1927/2011/1956$ and the strong variant's gaps are not larger in any matched slice (Appendix Figure 3).

Real-signal matched-IPS diagnostic. Figure 2 applies the same mechanics to Aotizhongxin PM2.5 from Beijing air quality data: fixed raw length gives unequal achieved information, pilot IPS matching brings all three preprocessing regimes to the same effective budget, and the matched view reduces the chance of attributing a gap difference to preprocessing when it is partly a budget imbalance. Smoothing changes the forecasting task, so this is a practice-facing check of allocate-and-verify on a public real signal, not a causal test of dependence strength.

Matching at a scarce real-signal budget. The same panel matched at the scarce target $N_{\text{eff}} = 100$ demonstrates the full allocate-verify-adjust loop at deployment scale: a 5,000-point deployment-scale pilot allocates raw lengths 3704/4178/4676 (raw/smooth6/smooth12), all within 15% of the fixed $N = 4096$, and achieved budgets land within 2% of target (101.8/101.3/99.5). At this common scarce budget, matched gaps are 0.112 (raw) versus 0.042/0.050 (smoothed), so the fixed-N ordering persists there. Comparing the two matched targets reveals what fixed-N evaluation cannot: the raw regime is by far the most budget-sensitive (0.112 at budget ≈ 100 down to 0.016 at ≈ 270, while the smoothed arms stay flat), so *which preprocessing appears best depends on the common budget level*; a fixed-N experiment pins every regime to one arbitrary slice of this axis (Appendix C.4).

Pilot-stage mixing diagnostic. The protocol presumes a finite integrated autocorrelation time. A practical pilot-stage check re-estimates $\widehat{\tau}_{\text{int}}$ under an increasing maximum lag M: a plateau indicates a stable budget, while growth without plateau flags long memory, where the protocol does not apply. On AR(1) ($\rho = 0.6$) the estimate is flat (3.94 across $M = 100$ to 5000); for long-memory autoregressive fractionally integrated moving average (ARFIMA) processes it grows by $5\times$ ($d = 0.30$) and $15\times$ ($d = 0.45$) without plateauing (Appendix C.5).

6 DISCUSSION AND LIMITATIONS

Matched-N_{eff} evaluation diagnoses information-budget confounding. Fixed raw length compares regimes with very different effective information; matching changes the interpretation of train–test gaps: the effect concentrates at scarce budgets, survives exposure control, holds across TCN, LSTM, and causal Transformer predictors, and on the Beijing panel cross-preprocessing conclusions can depend on the budget level. The conservative TCN bound is a separate anchor-risk baseline clarifying worst-case scaling, not predicting realized gaps.

Reporting checklist. For a temporal benchmark whose arms differ in dependence (Section 1 box), always report: raw N per regime; pilot $\widehat{\tau}$ and maximum lag; achieved N_{eff} with across-seed uncertainty; seed count; the split specification; context window and batch size; update counts whenever raw budgets differ; and both comparison tables.

Limitations. The causal matched-budget evidence is synthetic; the real-signal panels demonstrate the protocol, not causation. The theorem is a conservative anchor-risk reference, not a full-gap guarantee. Matching equalizes information, not exposure or task difficulty. Exponential mixing is assumed; the plateau diagnostic flags strong violations. Future work: non-stationarity, model ranking, full-risk bounds.

AI use statement

In this work, we used generative AI tools for editing and polishing the manuscript text and for writing experiment-orchestration and analysis scripts. We have

not used generative AI tools for generating the research ideas, deriving or proving the theoretical results, or producing the reported experimental numbers, and AI-based data collection is not applicable to this work. All AI-assisted text was reviewed and edited by the authors, and all AI-assisted code was verified by the authors by rerunning the experimental pipeline and cross-checking every reported statistic against the raw result files. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Baptiste Abeles, Eugenio Clerico, and Gergely Neu. Generalization bounds for mixing processes via delayed online-to-pac conversions. *arXiv preprint arXiv:2406.12600*, 2024.

Pierre Alquier and Benjamin Guedj. Simpler pac-bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

Shaojie Bai. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Peter Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear vc dimension bounds for piecewise polynomial networks. *Advances in neural information processing systems*, 1998.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20 (63):1–17, 2019.

Christoph Bergmeir, Rob J Hyndman, and Bonsoo Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, 2018.

Richard C Bradley. *Introduction to Strong Mixing Conditions*. Kendrick Press, Heber City, Utah, 2007.

A Philip Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–292, 1984.

Paul Doukhan. Mixing. In *Mixing: Properties and Examples*, pages 15–23. Springer, 1995.

Tomer Galanti, Mengjia Xu, Liane Galanti, and Tomaso Poggio. Norm-based generalization bounds for sparse neural networks. *Advances in Neural Information Processing Systems*, 36:42482–42501, 2023.

Charles J Geyer. Practical markov chain monte carlo. *Statistical science*, pages 473–483, 1992.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

Florian Graf, Sebastian Zeng, Bastian Rieck, Marc Niethammer, and Roland Kwitt. On measuring excess capacity in neural networks. *Advances in Neural Information Processing Systems*, 35:10164–10178, 2022.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for time series prediction with non-stationary processes. In *International conference on algorithmic learning theory*, pages 260–274. Springer, 2014.

Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

Antoine Ledent, Waleed Mustafa, Yunwen Lei, and Marius Kloft. Norm-based generalisation bounds for deep multi-class convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8279–8287, 2021.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4):1748–1764, 2021.

Philip M Long and Hanie Sedghi. Generalization bounds for deep convolutional neural networks. *arXiv preprint arXiv:1905.12600*, 2019.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. *Proceedings of the 27th International Conference on Algorithmic Learning Theory*, pages 3–17, 2016.

Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine learning*, 39(1):5–34, 2000.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in neural information processing systems*, 21, 2008.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 4: 1–26, 2009.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pages 1376–1401. PMLR, 2015.

Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*, 2019.

Alan Sokal. Monte carlo methods in statistical mechanics: foundations and new algorithms. In *Functional integration: Basics and applications*, pages 131–192. Springer, 1997.

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Daniel S Wilks. *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Sheldon X Chen. Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, 2017.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**] Sections 3 and 4 state the setting, the mixing and boundedness assumptions, and the model class; the introduction states the protocol as a numbered algorithm.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**] The bound in Section 4 makes the sample-size (anchor-count) scaling explicit; the protocol's cost is one pilot estimation and one verification pass per regime.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**No**] Code is not included at submission time; all datasets are public and Appendix B.1 specifies the complete configuration needed for reimplementation.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**] Assumptions on mixing, boundedness, and Lipschitz loss are stated in Section 3 and referenced by the theorem.
   (b) Complete proofs of all theoretical results. [**Yes**] All proofs appear in the appendix (Omitted Proofs).
   (c) Clear explanations of any assumptions. [**Yes**] Each assumption is discussed where introduced, and the limitations section discusses the exponential-mixing assumption together with a practical violation diagnostic.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**No**] Code is not released at submission time; both real datasets are standard public benchmarks (UCI Beijing Multi-Site Air Quality; PhysioNet MIT-BIH), and Appendix B.1 gives the full experimental specification.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**] Section 5 and Appendix B.1 give splits, hyperparameters, stopping rules, and the fixed-update variants.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**] The gap metric and the paired strong-minus-weak contrast are defined in Section 5; all headline results carry paired-bootstrap 95% confidence intervals over seeds.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**] All runs execute on a single consumer workstation (Apple M4 Max); each run trains a small model in minutes (Appendix B.1).

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**] The Beijing Multi-Site Air Quality dataset and the MIT-BIH Arrhythmia Database are cited.
   (b) The license information of the assets, if applicable. [**Yes**] Both are standard public benchmarks distributed under the open licenses of their repositories (UCI Machine Learning Repository; PhysioNet).
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Not Applicable**] No new assets are released.
   (d) Information about consent from data providers/curators. [**Not Applicable**] Both datasets are publicly distributed by their curators for research use.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**] The data are environmental measurements and de-identified physiological recordings.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

Evaluating Temporal Models at Matched Information: Supplementary Materials

Symbol	Description
Data	
$z_t \in \mathbb{R}^n$	Raw time series at time t
q	Context window length (lag)
$x_t \in \mathbb{R}^{q \times n}$	Windowed input $(z_{t-q+1}, \ldots, z_t)$
$y_t \in \mathbb{R}^n$	Target (e.g., z_{t+1})
N	Raw sequence length
$m = N - q$	Number of windowed examples
N_{eff}	Effective sample size (information proxy)
Dependence	
$\beta(k)$	β-mixing coefficient at lag k
C_0, c_0	Exponential mixing constants: $\beta(k) \le C_0 e^{-c_0 k}$
d, B	Anchor spacing and anchor count $B = \lfloor m/(d+1) \rfloor$ in the blocking proof
Architecture	
D	Depth (number of convolutional layers)
p	Kernel size
$W^{(\ell)}$	Convolution weights at layer ℓ
$\|W^{(\ell)}\|_{2,1}$	Filter-group norm at layer ℓ
$M^{(\ell)}$	Layer-wise norm budget: $\|W^{(\ell)}\|_{2,1} \le M^{(\ell)}$
R	Product budget: $R = \prod_{\ell=1}^{D} M^{(\ell)}$
$\mathcal{F}_{D,p,R}$	TCN hypothesis class under norm control
Learning	
$\ell(\cdot, \cdot)$	Loss function
$\mathcal{L}(f)$	Population risk
$\widehat{\mathcal{L}}_m(f)$	Empirical risk on m dependent examples
$\mathfrak{R}_M(\mathcal{F})$	(i.i.d.) Rademacher complexity on M samples

Table 2: Notation used throughout the paper.

N_{eff}	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
500	749	1,166	2,000	4,500
1000	1,499	2,333	4,000	9,000
2000	2,999	4,666	8,000	18,000
4000	5,999	9,333	16,000	36,000

Table 3: Raw sequence lengths emitted by the seeded implementation for the matched AR(1) sweep. Minor one-sample differences from closed-form values come from integer flooring in the configuration-to-length conversion.

A NOTATION AND APPENDIX ROADMAP

Appendix B follows the empirical story in the main text. Appendix B.1 gives the AR(1) matched-information protocol and reproducibility details; Appendix B.2 reports the full AR(1) sweep; Appendix B.3 gives the IPS-matched AR(2) check; Appendix B.4 gives real-signal IPS diagnostics; and Appendix B.11 collects depth, bound, and anchor diagnostics. Appendix D contains the proof pipeline. Table 2 collects notation used in both parts.

B ADDITIONAL EXPERIMENTAL RESULTS

This section follows the main empirical sequence: AR(1) matched-information evidence, IPS validation and AR(2) non-oracle matching, real-signal diagnostics, and then theory-aligned and bound/anchor diagnostics. The controlled matched-N_{eff} claims rely on the seeded synthetic experiments; real-data plots are observational diagnostics.

Run counts and grids. The validation-split fair-comparison experiment uses 4 dependence levels, 4 target effective sizes, 4 depths, and 10 seeds, for $4 \times 4 \times 4 \times 10 = 640$ runs. Figure 1 reports the full sweep.

ρ	true τ_{int}	IPS $\widehat{\tau}_{\text{int}}$	relative error
0.20	1.50	1.49	-0.7%
0.40	2.33	2.31	-1.1%
0.60	4.00	3.93	-1.7%
0.80	9.00	8.93	-0.8%
0.90	19.00	18.88	-0.6%
0.95	39.00	38.86	-0.4%

Table 4: IPS validation on long AR(1) sequences. The estimator recovers the oracle integrated autocorrelation time to within 2% across dependence levels, supporting its use as an operational N_{eff} proxy.

B.1 AR(1) matched-information protocol and reproducibility

Figure 1 and Table 7 use the full 640-run sweep; the depth/anchor and conservative-bound diagnostics use the same seeded protocol restricted to the $N_{\text{eff}} = 2000$ slice. The synthetic process is $z_t = \rho z_{t-1} + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}(0, 0.5^2)$; we discard a burn-in of 200 raw samples, and all reported raw lengths N exclude this burn-in. Each raw sequence is split chronologically into 64% train, 16% validation, and 20% test segments. We then construct windows only within each split, fit normalization statistics on the training segment, and use the validation segment for early stopping.

The $N_{\text{eff}} = 2000$ slice used synthetic data with fair-comparison matching enabled, depths $D \in \{2, 4, 6, 8\}$, ten trials per depth, and eight worker processes. Results were stored in the BatchNorm $N_{\text{eff}} = 2000$ result directory.

ρ	raw N	train	val.	test	$\widehat{N}_{\text{eff}}^{\text{IPS}}$	anchors	delay	gap
0.2	2,999	1,887	448	568	1922 ± 35	50	37	-0.010 ± 0.012
0.4	4,666	2,954	714	902	1958 ± 20	71	41	0.012 ± 0.008
0.6	8,000	5,088	1,248	1,568	2029 ± 10	102	49	-0.008 ± 0.006
0.8	18,000	11,488	2,848	3,568	2029 ± 18	154	74	-0.002 ± 0.006

Table 5: Exact lengths and aggregate diagnostics for the seeded validation-split experiment. Train/validation/test columns count supervised windows after within-split construction. Gaps and IPS estimates are mean ± s.e. over 40 trial-depth cells. Anchor counts are computed on the training-split windows, $B \approx m_{\text{train}}/(d+1)$, with the window-corrected delay convention.

D	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
2	-0.053 ± 0.024	0.016 ± 0.017	-0.022 ± 0.011	-0.006 ± 0.017
4	0.010 ± 0.029	0.014 ± 0.021	0.003 ± 0.008	0.002 ± 0.004
6	-0.034 ± 0.022	0.013 ± 0.014	-0.004 ± 0.011	-0.002 ± 0.008
8	0.035 ± 0.013	0.004 ± 0.015	-0.010 ± 0.014	-0.003 ± 0.014

Table 6: Depth-stratified generalization gaps for the seeded experiment, reported as mean ± s.e. over ten trials per cell. The main text aggregates these values descriptively over depths; this table exposes the between-depth heterogeneity.

The BatchNorm sweep for Figure 1 ran on an Apple M4 Max workstation (16 CPU cores, 64 GB unified memory). The $N_{\text{eff}} = 2000$ slice used 8 CPU workers, took about 22 minutes wall time, and accounts for 2.1 cumulative training CPU-hours (median 42 seconds per run, maximum 322 seconds); the additional $N_{\text{eff}} \in \{500, 1000, 4000\}$ slices required about 18.7 cumulative CPU-hours and 3 wall-clock hours. The theory-aligned ablation in Appendix B.5 ran on an institutional Slurm CPU cluster with 32 workers, 64 GB memory, and 18:52 wall time. Local runs used runtime 3.13.9, PyTorch 2.10.0, NumPy 2.3.5, pandas 2.3.3, SciPy 1.16.3, Matplotlib 3.10.6, WFDB 4.3.1, and PyYAML 6.0.3; the cluster ablation used runtime 3.10.20, PyTorch 2.6.0, NumPy 2.2.6, pandas 2.3.3, SciPy 1.15.3, and PyYAML 6.0.3. Each CSV row stores the deterministic seed for its $(N_{\text{eff}}, \rho, D, \text{trial})$ tuple.

B.2 Full AR(1) matched-information sweep

The main fair-comparison figure uses the same chronological train/validation/test protocol for target effective sizes $N_{\text{eff}} \in \{500, 1000, 2000, 4000\}$, with the same four dependence levels, four depths, and ten trials per depth. This yields 640 completed runs with 40 trial-depth cells per (N_{eff}, ρ) point. The sweep used synthetic data with fair-comparison matching enabled, depths $D \in \{2, 4, 6, 8\}$, ten trials per depth, and eight worker processes. The expected output has 640 rows, with 10 rows for each $(N_{\text{eff}}, \rho, D)$ cell. Figure 1 uses this CSV and the same plotting utilities as the appendix diagnostics.

N_{eff}	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
500	0.137 ± 0.014	0.088 ± 0.012	0.081 ± 0.012	0.074 ± 0.009
1000	0.064 ± 0.009	0.064 ± 0.009	0.047 ± 0.006	0.030 ± 0.005
2000	0.060 ± 0.008	0.043 ± 0.005	0.031 ± 0.003	0.024 ± 0.004
4000	0.038 ± 0.005	0.021 ± 0.003	0.021 ± 0.004	0.021 ± 0.003

Table 7: Mean absolute empirical gaps in the 640-run matched-information sweep, reported as mean ± s.e. over 40 trial-depth cells per condition. At every target N_{eff}, the strongest dependence setting does not have a larger mean absolute gap than the weakest setting.

For the paired trial-depth contrast $\rho = 0.8$ minus $\rho = 0.2$, the mean absolute-gap differences are -0.063 ± 0.017, -0.034 ± 0.010, -0.037 ± 0.009, and -0.017 ± 0.006 for $N_{\text{eff}} = 500, 1000, 2000, 4000$ (approximate 95% CIs $[-0.096, -0.030]$, $[-0.054, -0.014]$, $[-0.055, -0.019]$, and $[-0.029, -0.005]$). The main-text monotone-trend check fits, within each matched target-depth-trial cell, the slope of |Gap| versus ρ across the four dependence levels and bootstraps the 160 paired slopes. These are diagnostics: once effective information is matched, dependence strength alone is not a reliable proxy for gap size. The aggregate pattern is not driven by a single depth: the $\rho = 0.8$ minus $\rho = 0.2$ absolute-gap difference is negative at all four N_{eff} slices for $D \in \{4, 6, 8\}$, and for $D = 2$ it is negative in three of four slices (the exception is the smallest $N_{\text{eff}} = 500$ slice, where the difference is 0.016).

B.3 AR(2) IPS-matched positive-persistence check and fixed-N confounding

The AR(1) sweep is useful because the oracle effective sample size is known, but it does not test non-oracle matching. We therefore evaluate a second controlled family of stationary AR(2) processes,

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, 0.5^2),$$

using stable positive-dependence variants: weak $(0.25, 0.05)$, medium $(0.55, 0.10)$, and strong $(0.75, 0.10)$. These coefficients satisfy the standard AR(2) stability conditions; with Gaussian innovations, the processes are geometrically mixing. Raw lengths come from finite pilot IPS estimates rather than an analytic AR(1) formula. On length-100,000 pilot sequences, $\widehat{\tau}_{\text{int}} = 1.904, 5.112, 13.142$ for weak, medium, and strong AR(2), so runs use $N = \lceil N_{\text{eff}} \widehat{\tau}_{\text{int}} \rceil$. Matching is approximate; Table 8 reports achieved IPS effective sizes within about 5% of target.

The matched AR(2) sweep uses $N_{\text{eff}} \in \{1000, 2000, 4000\}$, the same depths $D \in \{2, 4, 6, 8\}$, and six trials per cell, for $3 \times 3 \times 4 \times 6 = 216$ runs. The fixed-N comparison uses the same three AR(2) variants, $N = 4096$, four depths, and six trials, for 72 runs. Both were run on an institutional Slurm CPU cluster with 24 worker processes and 16:53 wall-clock time. Pilot IPS estimates used length-100,000 pilot sequences with maximum lag 500.

Figure 3 visualizes both sides of the AR(2) check: IPS matching brings the achieved effective sizes close to target, and the strong-dependence variant is not larger in absolute gap after matching.

B.4 Real-signal IPS diagnostics

Real sequences play two roles here. Table 10 reports non-matched IPS diagnostics: PhysioNet ECG segments have $\widehat{\rho} \in \{0.935, 0.958\}$ and $\widehat{N}_{\text{eff}}/N \approx 3$–$4\%$, while Beijing Multi-Site Air Quality variables (Zhang et al., 2017) show a 178× spread in estimated effective-sample fraction between slowly varying pollutants and rainfall. Figure 2 in the main text gives a small matched-IPS protocol demonstration on one Beijing PM2.5 signal. Together, these diagnostics show that information content can vary sharply at fixed raw length.

Beijing matched-IPS protocol demonstration. We use the Aotizhongxin PM2.5 series from the Beijing dataset and compare the raw hourly signal with trailing 6-hour and 12-hour moving averages. Missing values are

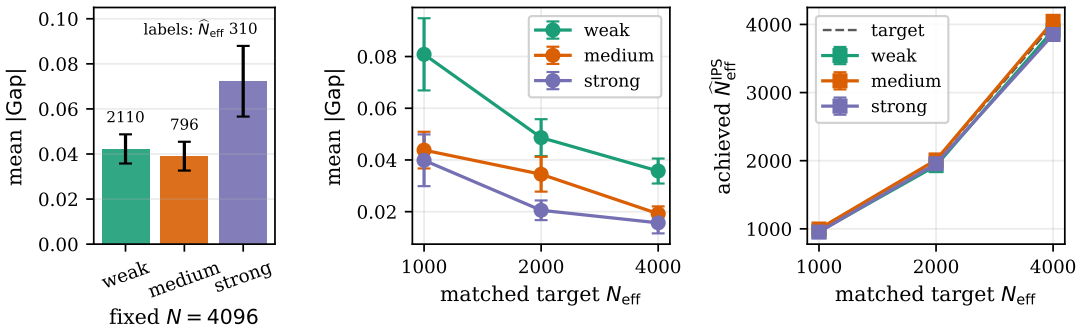


Figure 3: **AR(2) positive-persistence check using IPS-estimated matching.** At fixed $N = 4096$, achieved $\widehat{N}_{\text{eff}}$ differs sharply across variants; after pilot IPS matching, achieved effective sizes track the targets and gaps do not increase for the strong variant. Points aggregate 24 trial-depth cells.

target N_{eff}	AR(2) variant	raw N	achieved $\widehat{N}_{\text{eff}}^{\text{IPS}}$	|Gap|	signed gap
1000	weak	1,904	961 ± 25	0.081 ± 0.014	-0.012 ± 0.022
1000	medium	5,112	997 ± 16	0.044 ± 0.007	-0.001 ± 0.012
1000	strong	13,142	952 ± 19	0.040 ± 0.010	0.014 ± 0.013
2000	weak	3,808	1927 ± 35	0.049 ± 0.007	0.004 ± 0.012
2000	medium	10,224	2011 ± 30	0.034 ± 0.007	-0.002 ± 0.010
2000	strong	26,284	1956 ± 38	0.021 ± 0.004	0.003 ± 0.006
4000	weak	7,616	3914 ± 55	0.036 ± 0.005	-0.007 ± 0.009
4000	medium	20,448	4046 ± 45	0.019 ± 0.003	0.001 ± 0.005
4000	strong	52,568	3853 ± 41	0.016 ± 0.004	0.012 ± 0.005

Table 8: AR(2) matched-information positive-persistence check. Values are mean ± s.e. over 24 trial-depth cells per row. Paired strong-minus-weak absolute-gap contrasts are -0.041 ± 0.016, -0.028 ± 0.009, and -0.020 ± 0.006 for target $N_{\text{eff}} = 1000, 2000, 4000$, with approximate 95% CIs $[-0.072, -0.010]$, $[-0.045, -0.011]$, and $[-0.031, -0.009]$.

interpolated before constructing chronological splits. Because smoothing changes the task and the longer matched windows can change the calendar distribution, this is an operational protocol check rather than a causal test of dependence strength. We estimate IPS effective size from a pilot prefix, choose raw lengths to approximately match a target effective size, and verify achieved effective sizes and train–test gaps.

The fixed-raw-N absolute gaps are 0.098 ± 0.012, 0.041 ± 0.020, and 0.025 ± 0.011 for raw, smooth6, and smooth12. After pilot IPS matching they are 0.016 ± 0.004, 0.027 ± 0.005, and 0.021 ± 0.004. These numbers are an end-to-end protocol check on a public real signal, not evidence that dependence alone causes the changes.

Dataset access and licenses. The MIT-BIH Arrhythmia Database is publicly available through PhysioNet (version 1.0.0, DOI `https://doi.org/10.13026/C2F305`) under the Open Data Commons Attribution License v1.0. The Beijing Multi-Site Air Quality dataset is available through UCI (DOI `https://doi.org/10.24432/C5RK5G`) under CC BY 4.0. We use both as public time-series diagnostics and cite the dataset papers in the main text.

B.5 Theory-aligned no-BatchNorm ablation

To reduce the gap between the formal TCN class and the trained architecture, we also evaluate a theory-aligned 640-run matched-information sweep with the same target effective sizes and depths as the main multi-slice sweep, ten trials per cell, 32 worker processes, hidden-layer BatchNorm disabled, output clipping at 5.0, and weight projection enabled throughout training. This ablation is still not a calibrated theorem check (the Gaussian responses are not almost surely bounded), but it removes the most visible architectural mismatch: hidden-layer BatchNorm.

The paired trial-depth contrasts $|\text{Gap}|_{\rho=0.8} - |\text{Gap}|_{\rho=0.2}$ are -0.020 ± 0.021, -0.013 ± 0.012, -0.020 ± 0.009, and -0.006 ± 0.006 for $N_{\text{eff}} = 500, 1000, 2000, 4000$, respectively. Signed gaps remain centered near zero across

raw N	AR(2) variant	achieved $\widehat{N}_{\text{eff}}^{\text{IPS}}$	|Gap|	signed gap
4,096	weak	2110 ± 31	0.042 ± 0.006	-0.003 ± 0.011
4,096	medium	796 ± 21	0.039 ± 0.006	0.009 ± 0.010
4,096	strong	310 ± 10	0.072 ± 0.016	0.019 ± 0.021

Table 9: Fixed-raw-N AR(2) comparison. Holding $N = 4096$ fixed changes the achieved effective information budget by about 6.8× between weak and strong AR(2). The strong variant has the largest numerical absolute gap, but the paired strong-minus-weak contrast is 0.030 ± 0.018 (approximate 95% CI $[-0.006, 0.066]$), so this panel is best read as an information-imbalance diagnostic.

source	IPS diagnostic	interpretation
PhysioNet ECG	$\widehat{\rho} \in \{0.935, 0.958\}$; $\widehat{N}_{\text{eff}}/N \approx 3$–$4\%$	strong serial dependence
Beijing air quality	178× spread across variables	same raw N, very different information
Oscillatory AR(2) sanity check	$(\phi_1, \phi_2) = (0.60, -0.45)$; $\widehat{\tau}_{\text{int}} = 1.83$	IPS truncates after negative paired ACF

Table 10: IPS diagnostics beyond the oracle AR(1) setting. The oscillatory AR(2) row is an estimator stress check only (not a training sweep): on a length-100,000 pilot path, the lag-1 and lag-2 ACFs are 0.415 and −0.202, giving $\widehat{N}_{\text{eff}}/N \approx 0.546$.

the same grid. Thus the stronger-dependence regime no longer shows as large an absolute-gap reduction as in the BatchNorm runs, but it still does not produce larger gaps than the weak-dependence regime once effective information is matched.

B.6 Synthetic data: optimal delay parameter analysis

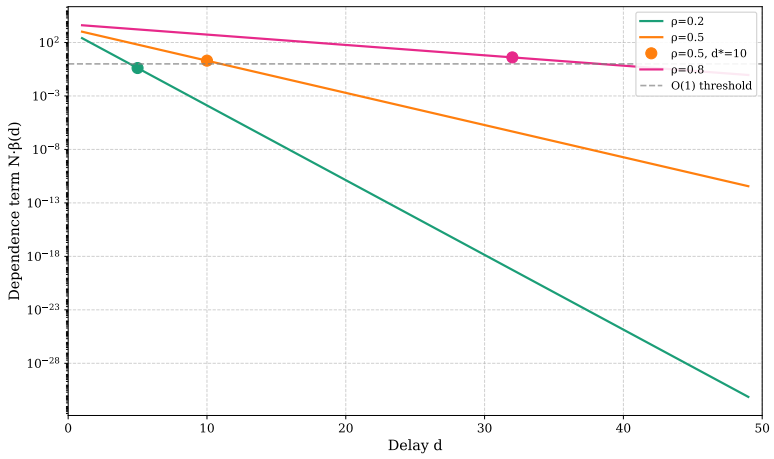


Figure 4: **Effect of delay d on the mixing-dependent penalty at $N = 16{,}384$.** The proxy $N \cdot \beta(d)$ shows how larger delay reduces residual dependence under exponential β-mixing. In Theorem 4.2, dependence enters through $(B-1)\beta(d+1)$ with $B = \lfloor m/(d+1) \rfloor$ for the windowed example process, so $N\beta(d)$ is a conservative raw-process proxy. The orange marker shows the raw-process choice $d^* = \lceil \ln N/c_0 \rceil$; for windowed examples add the fixed context offset in Remark 3.2.

Section 4 motivates choosing the delay parameter $d^* = \lceil \ln m/c_0 \rceil$ for the windowed example process when $\beta(k) \le C_0 e^{-c_0 k}$. This balances (i) reducing dependence via $\beta(d)$ with (ii) preserving enough effective anchors $B = \lfloor m/(d+1) \rfloor$. Figure 4 visualizes how the raw-process proxy $N\beta(d)$ decays with d across dependence regimes. If one starts from raw AR(1) coefficients rather than the already-windowed process, Remark 3.2 gives the corresponding q-offset.

Relating ρ to β-mixing (Gaussian AR(1)). Consider the stationary Gaussian AR(1) process $Z_t = \rho Z_{t-1} + \varepsilon_t$ with $|\rho| < 1$ and i.i.d. Gaussian noise (ε_t). This process is *geometrically* β-mixing: there exist constants $C(\rho) \ge 1$

N_{eff}	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
500	0.134 ± 0.014	0.098 ± 0.013	0.104 ± 0.013	0.114 ± 0.013
1000	0.065 ± 0.009	0.072 ± 0.011	0.064 ± 0.009	0.052 ± 0.007
2000	0.064 ± 0.008	0.049 ± 0.005	0.043 ± 0.004	0.043 ± 0.006
4000	0.037 ± 0.005	0.026 ± 0.003	0.033 ± 0.004	0.031 ± 0.004

Table 11: Mean absolute empirical gaps in the theory-aligned ablation (no BatchNorm, output clipping at 5.0, and projected convolutional weights), reported as mean ± s.e. over 40 trial-depth cells per condition. Removing BatchNorm weakens the contrast but preserves the main diagnostic pattern: at matched information, the strongest dependence setting is not larger than the weakest dependence setting.

such that

$$\beta(k) \;\le\; C(\rho)\,|\rho|^{k} \;=\; C(\rho)\,e^{-k\cdot(-\log|\rho|)}.$$

Hence it satisfies Assumption 3.1 with rate parameter $c_0 \asymp -\log|\rho|$ (and $C_0 = C(\rho)$ absorbed into constants). Therefore, larger $\rho \uparrow 1$ implies smaller c_0 and slower mixing, so a larger delay d is required to make $\beta(d+1)$ small. (Bradley, 2007; Doukhan, 1995)

Autocorrelation-based effective sample size under AR(1). If one wishes to connect ρ to the classical *ACF-based* notion of effective sample size, then for AR(1) the integrated autocorrelation time is $\tau_{\text{int}} = 1 + 2\sum_{k\ge1}\rho^k = \frac{1+\rho}{1-\rho}$, giving

$$N_{\text{eff}}^{(\text{ACF})} \;=\; \frac{N}{\tau_{\text{int}}} \;=\; N\cdot\frac{1-\rho}{1+\rho}.$$

(We emphasize this is distinct from the anchor count $B = \lfloor m/(d+1)\rfloor$, which is tied to β-mixing via blocking.) (Geyer, 1992; Sokal, 1997)

B.7 Synthetic data: weight norm behavior

The theoretical baseline in Section 4 depends on layer-wise norm control: if $\|W^{(\ell)}\|_{2,1} \le M^{(\ell)}$, the bound scales with $R = \prod_{\ell=1}^{D} M^{(\ell)}$. In experiments, we monitor an empirical aggregate filter-norm proxy from the learned weights. The implementation-level projection and weight decay encourage smaller norms, but they are not identical to the per-layer product constraint in the theorem, so these plots diagnose how optimization and regularization interact with raw sequence length.

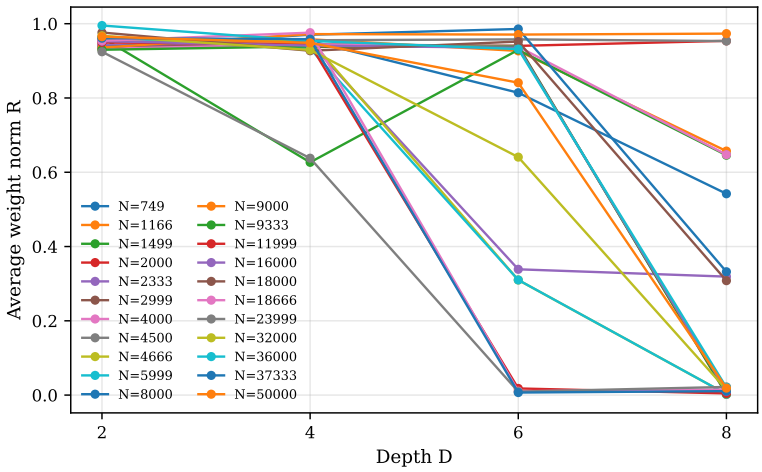


Figure 5: **Weight norm versus depth for different raw sequence lengths (synthetic).** The empirical norm proxy is computed consistently across runs. Long sequences (e.g., $N = 16{,}384$) can induce larger norms than shorter sequences, suggesting that raw length can affect regularization needs even when effective information is matched. Norms also tend to increase with depth.

Figure 5 shows empirical norms increasing with depth across raw sequence lengths. The dependence on N differs across regimes, motivating the main-text practice of reporting both N and N_{eff}.

B.8 Physiological data (PhysioNet): empirical scaling on real signals

We also evaluate physiological ECG data (PhysioNet) as a real-signal diagnostic. Because ECG dependence is unknown and not directly controllable, these fixed-raw-N experiments probe scaling with sequence length and depth; they are not evidence for the controlled matched-N_{eff} phenomenon.

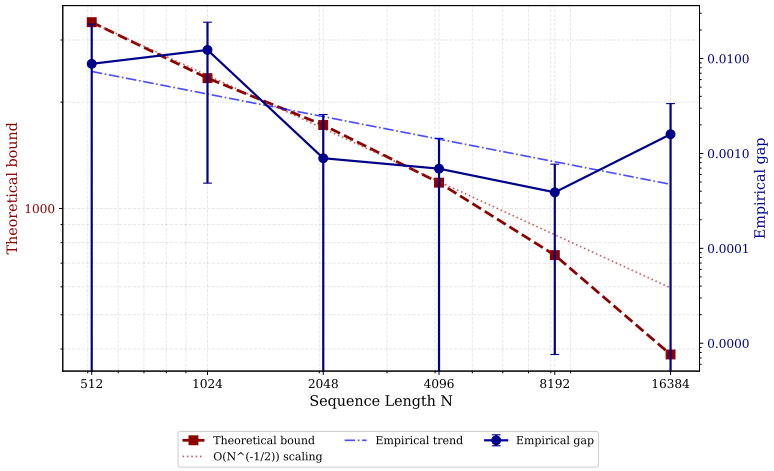


Figure 6: **PhysioNet: empirical generalization gap and theoretical bound vs. sequence length.** The empirical gap (blue; right y-axis) decreases with N in this fixed-raw-N diagnostic. The theoretical bound (red; left y-axis) also decreases but remains far above the measured gaps.

Sequence-length scaling. Figure 6 shows the empirical gap decreasing with N. This suggests structured regularities (e.g., quasi-periodicity and constrained dynamics) beyond the generic worst-case dependent-process baseline.

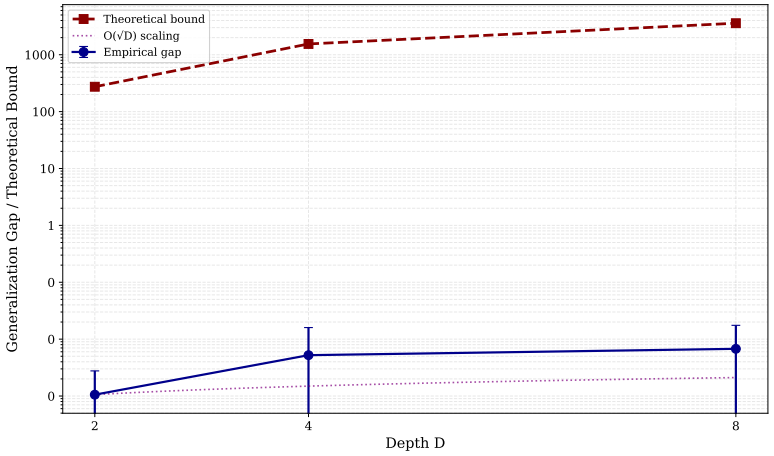


Figure 7: **PhysioNet: empirical generalization gap and theoretical bound vs. depth.** The empirical gaps (blue) increase with depth in this experiment, whereas our norm-controlled baseline suggests milder $O(\sqrt{D})$ dependence in the architectural complexity term. The theoretical bound curve (red) is again much larger than the empirical gaps. Error bars show ± 1 s.e. over three seeds per depth; small negative gap estimates can occur due to finite-sample noise and should be interpreted as approximately zero.

Depth scaling. Figure 7 indicates that empirical gaps grow with depth on this dataset. This is plausible under optimization and finite-sample effects; the theoretical $\sqrt{D}$ term is worst-case architecture dependence, not a

prediction of realized training dynamics.

Overall, the PhysioNet plots show bound conservatism and dataset-specific structure in uncontrolled real signals.

B.9 PhysioNet Weight Norm Dynamics

We next examine the weight-norm dynamics on physiological ECG data (PhysioNet). Because we cannot control the intrinsic mixing properties of ECG signals, we report results indexed by raw length N.

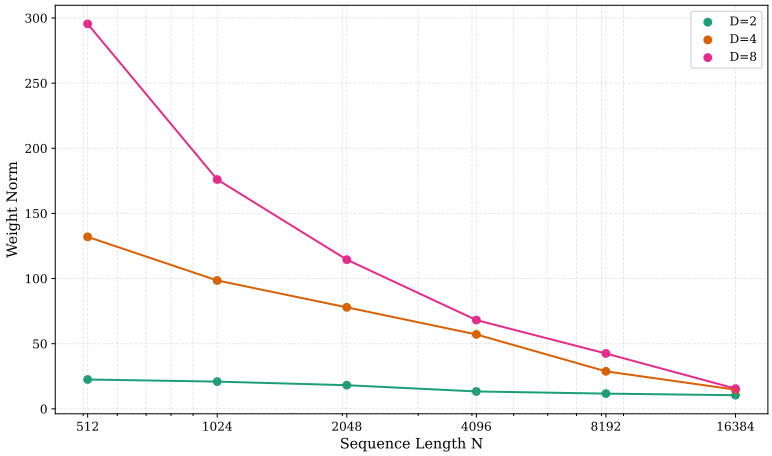


Figure 8: Empirical norm proxy versus raw sequence length on PhysioNet. We use raw N because the mixing properties are not controllable. The steepest decline occurs between $N = 512$ and $N = 2048$.

Figure 8 shows an inverse relationship between the empirical norm proxy and N, in contrast to some synthetic regimes.

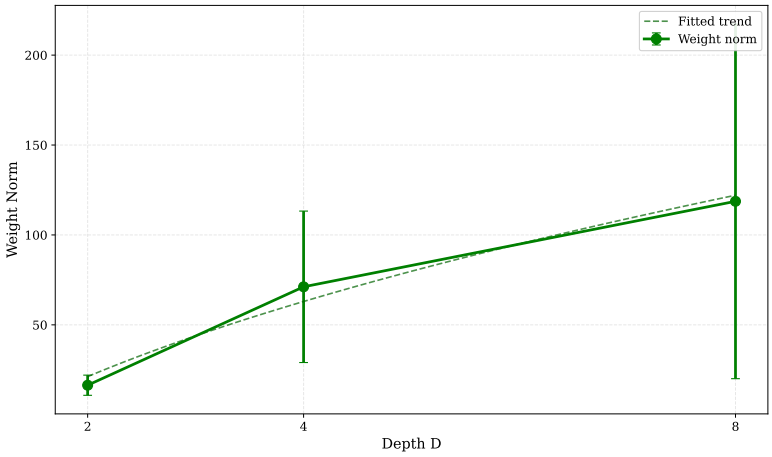


Figure 9: **PhysioNet: norm growth with depth.** A fitted linear trend (dashed line) indicates approximately linear growth in the empirical norm proxy as layers are added.

Figure 9 indicates that the reported norm proxy grows approximately linearly with depth in this dataset. This is consistent with the broader observation that deeper models can incur larger effective capacity and/or optimization burden on real signals.

B.10 Additional PhysioNet Architecture Diagnostic

Figure 10 suggests that intermediate depth ($D = 4$) can behave differently with N in this dataset. Because these are fixed-raw-N experiments with uncontrolled ECG dependence, architecture-dependent trends may reflect

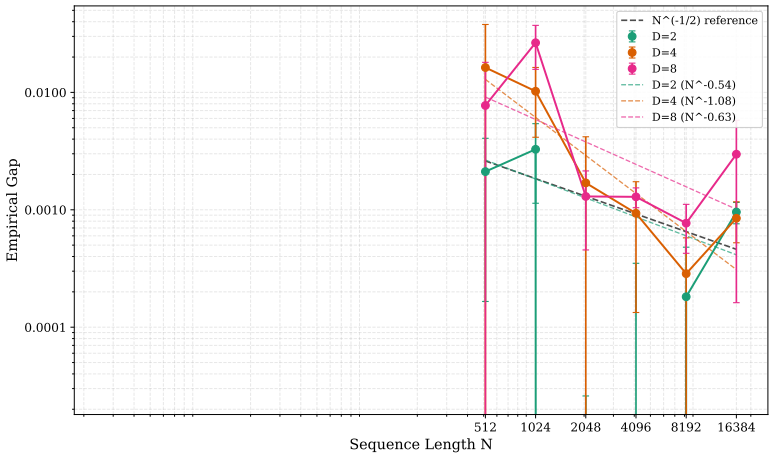


Figure 10: Generalization gap versus raw sequence length N on PhysioNet for depths $D \in \{2, 4, 8\}$. Lines show fitted power-law exponents; error bars denote ± 1 s.e. over three runs.

dataset-specific information and optimization dynamics.

B.11 Depth, bound, and anchor diagnostics for matched AR(1)

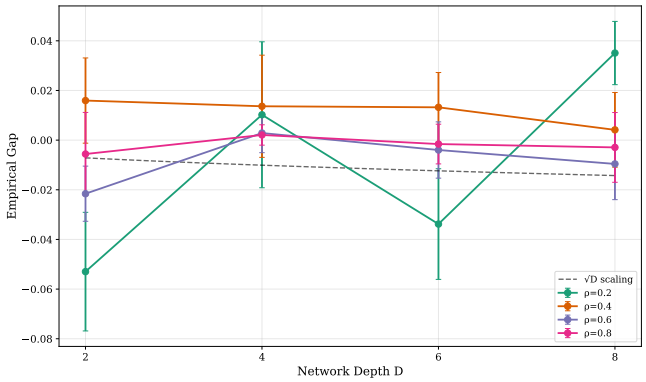


Figure 11: **Depth scaling under fair comparison.** Empirical gaps at $N_{\text{eff}} = 2000$ for depths $D \in \{2, 4, 6, 8\}$ and dependence strengths $\rho \in \{0.2, 0.4, 0.6, 0.8\}$. The dashed line shows a $\sqrt{D}$ reference trend. Error bars represent ± 1 standard error over ten seeds.

Figure 11 is a depth diagnostic rather than a new claim: it checks whether the matched-information pattern is concentrated at one depth or persists across the tested TCN depths.

Figure 12 shows that, under matched N_{eff}, the conservatism level of the theoretical diagnostic is broadly consistent across dependence strengths and depths. This supports its intended role as a loose worst-case reference.

Figures 13 and 14 explain the empirical side of the anchor/full-risk limitation. The anchor estimator is visibly noisier than the full training risk, while block averaging gives a computable retained estimator that can move closer to the full risk without changing the block-level stochastic unit.

B.12 Architectural-factor sanity check against black-box dependence-aware bounds

Using Mohri–Rostamizadeh as a black box controls dependence but leaves the i.i.d. complexity term architecture-agnostic unless one plugs in a neural-network bound. In our synthetic setting ($D = 4$, $p = 3$, $C = 32$, $q = 32$), a crude parameter-count proxy has about 9.3×10^3 weights and contributes $\sqrt{P/B} \approx 9.6$ at representative anchor count $B \approx 102$. The convolution-aware term in Theorem 4.4 contributes $\sqrt{D \log(2p)/B} \approx 0.27$ before Λ_{TCN},

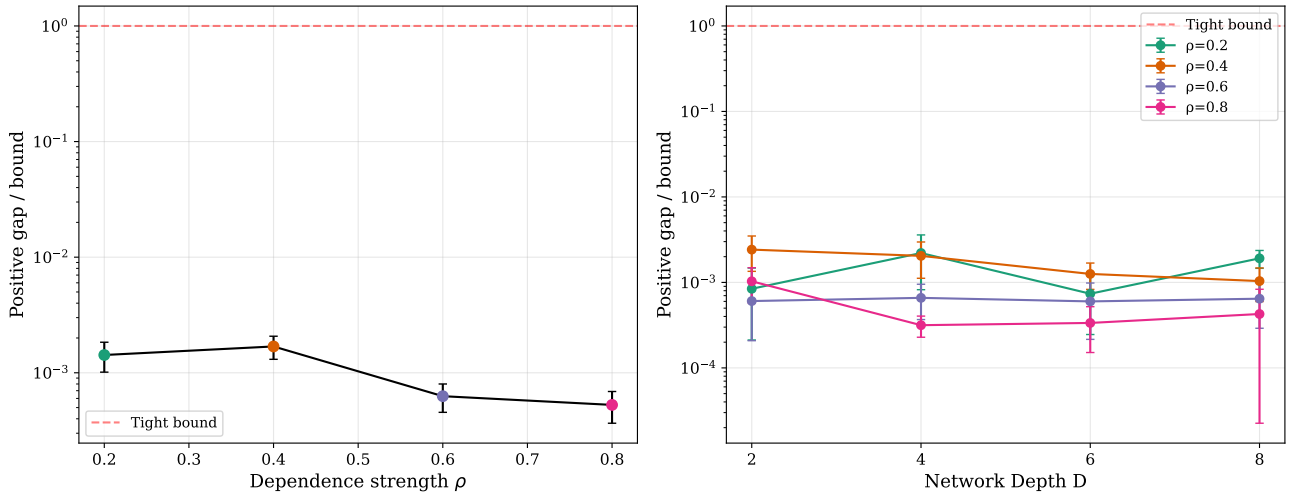


Figure 12: **Conservative bound diagnostic versus empirical gaps.** We plot $\max\{\text{Gap}, 0\}/(\text{theoretical diagnostic})$ at matched $N_{\text{eff}} = 2000$. Values near 0 mean the conservative bound is much larger than the measured train–test gap; values closer to 1 would indicate tightness. The ratio remains well below 1, supporting the intended role of the theory as a loose worst-case reference rather than a calibrated predictor.

norm products, input bounds, and loss constants. This is not a bound-value comparison; it only shows that the CNN ingredient replaces generic parameter-count dependence with a convolution-aware architectural factor.

B.13 Empirical calibration of the bound constants

We do not estimate a universal constant for Theorem 4.4. On the 160-run validation-split slice, a diagnostic regression of signed gap on the empirical norm proxy times $\sqrt{D\log(2p)\log N/N}$ has low explanatory power ($R^2 \approx 0.01$), consistent with a worst-case baseline rather than a calibrated predictor.

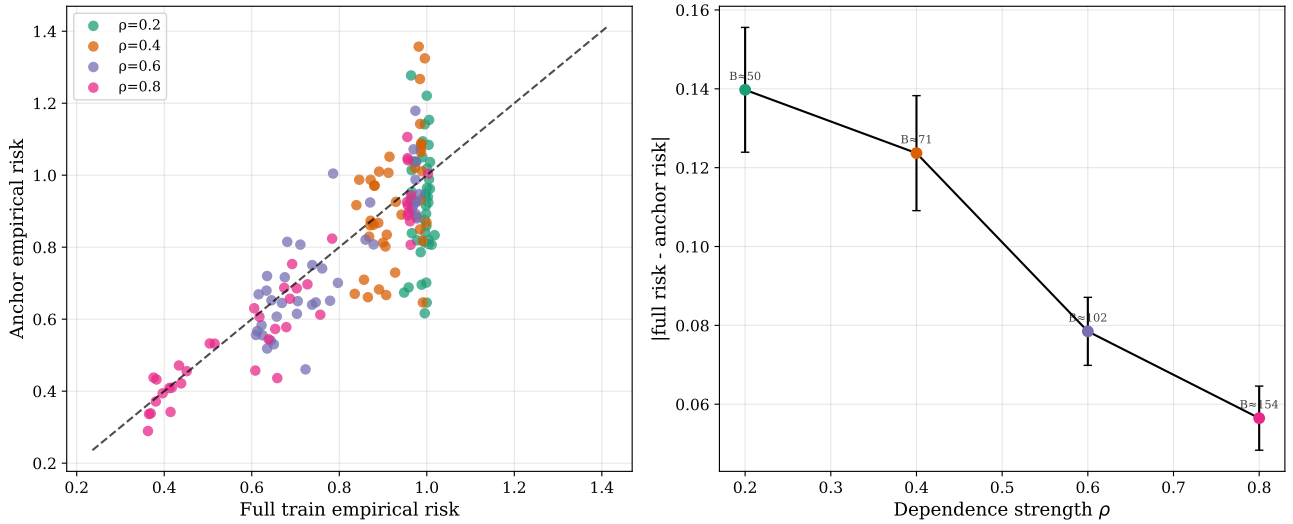


Figure 13: **Anchor empirical risk versus full training empirical risk** at $N_{\text{eff}} = 2000$. Left: anchor risk versus full empirical train risk; the dashed line is equality. Right: mean absolute discrepancy by dependence strength, with average anchor counts annotated. The anchor estimator is noisier than the full risk because it uses one point per block; discrepancies are smaller in high-dependence matched-information regimes where raw sequences and anchor counts are larger.

C SCARCE-BUDGET, EXPOSURE-CONTROL, AND GENERALITY EXPERIMENTS

This section reports the matched-protocol experiments at scarce information budgets, the fixed-update exposure controls, the architecture-generality runs, the Beijing panel matched at a scarce budget, and the long-memory pilot diagnostic referenced from Section 5. All runs use the recipe of Section 5 unless stated otherwise, and all contrasts are strong-minus-weak |Gap| contrasts ($\rho = 0.8$ minus $\rho = 0.2$), paired by depth and trial, with paired-bootstrap 95% confidence intervals.

C.1 Scarce-budget AR(1) sweep

The sweep covers targets $N_{\text{eff}} \in \{50, 100, 200\}$, $\rho \in \{0.2, 0.4, 0.6, 0.8\}$, depths $D \in \{2, 4\}$, and 20 trials per cell, in four stages (960 runs): context window $q = 16$ at targets $\{100, 200\}$ and $q = 8$ at target 50, each with and without early stopping. The window is shortened, constant across ρ within each target, because the weak-dependence arm at $N_{\text{eff}} = 50$ has only 75 raw points, which is window-degenerate at $q = 32$. The no-early-stopping stages set the patience above the epoch cap so every run trains for exactly the full 100-epoch budget.

The regime is genuinely unsaturated: mean |Gap| spans 0.12 to 0.36, and signed gaps at weak dependence are strongly positive (at $N_{\text{eff}} = 100$, $\rho = 0.4$: mean signed gap +0.206 with 65% of runs positive), i.e., systematic overfitting rather than noise around zero. Table 12 reports the per-cell magnitudes and contrasts. At $N_{\text{eff}} = 100$, |Gap| decreases monotonically in ρ and the contrast is largest; at $N_{\text{eff}} = 50$ the profile peaks at mid ρ while the strong arm stays below the weak arm, and the contrasts remain non-positive with confidence intervals that include zero.

C.2 Fixed-update exposure control

Matched regimes have different raw lengths, so a fixed epoch budget implies different numbers of gradient steps per regime. The fixed-update variant performs exactly U Adam updates (batch size 64, as in all experiments) in every run regardless of ρ, with no validation-based stopping and no best-checkpoint selection; the final weights are evaluated. Three grids of 160 runs each (20 trials $\times$ depths $\{2, 4\}$ $\times$ 4 values of ρ):

- $U = 200$ at target $N_{\text{eff}} = 100$ ($q = 16$; raw lengths 150/233/400/900): |Gap| 0.360/0.257/0.256/0.159 for

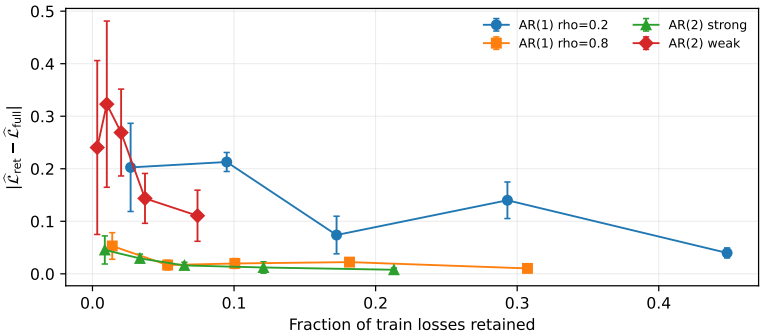


Figure 14: **Block-compatible empirical risk sanity check.** For four matched synthetic conditions, we compare the full training risk with retained estimators. Each curve moves from one-point anchors to block averages of length $a \in \{4, 8, 16, 32\}$ using the same gap length as the anchor delay; points farther right retain more training losses. Block averaging can track full training risk more closely, while the stochastic unit remains the retained block. Small cells are noisy, so we do not claim monotone improvement for every condition.

Table 12: Scarce-budget AR(1) sweep: mean |Gap| per ρ (early-stopping stages) and paired contrasts. Each cell aggregates 40 runs (20 trials $\times$ depths $\{2, 4\}$).

	|Gap| by ρ (ES)				Contrast ($\rho = 0.8$ minus $\rho = 0.2$)		
N_{eff}	0.2	0.4	0.6	0.8	Early stop	No early stop	95% CI (ES)
50 ($q = 8$)	0.297	0.343	0.328	0.261	-0.036	-0.052	includes 0
100 ($q = 16$)	0.354	0.343	0.256	0.152	-0.203 ± 0.060	-0.210	$[-0.321, -0.085]$
200 ($q = 16$)	0.215	0.165	0.115	0.119	-0.096 ± 0.036	-0.096	$[-0.166, -0.026]$

$\rho = 0.2/0.4/0.6/0.8$; contrast -0.201 ± 0.048 (CI $[-0.294, -0.107]$), unchanged from the standard recipe's -0.203.

- $U = 2000$ at target $N_{\text{eff}} = 2000$ ($q = 32$): |Gap| 0.061/0.056/0.055/0.074; contrast $+0.013 \pm 0.025$ (CI $[-0.035, +0.062]$).
- $U = 4000$ at target $N_{\text{eff}} = 2000$: |Gap| 0.063/0.041/0.073/0.076; contrast $+0.013 \pm 0.018$ (CI $[-0.022, +0.048]$), an identical point estimate, so the rich-budget null is robust to the update-budget choice.

Read jointly with the main sweeps: the no-increase conclusion holds under fixed exposure at every budget. At scarce budgets the negative contrast is exposure-invariant, hence information-driven; at rich budgets the standard recipe's residual negative contrast (-0.037 at $N_{\text{eff}} = 2000$) attenuates to a null under fixed exposure, so part of that residual reflects training duration. The protocol therefore prescribes reporting update counts whenever raw budgets differ across regimes.

C.3 Architecture generality: LSTM and causal Transformer

We rerun the matched protocol with a 2-layer LSTM and a 2-layer causal Transformer ($d_{\text{model}} = 32$, 2 heads), 10 seeds per cell, same optimizer, chronological splits, and windowing recipe, at targets $N_{\text{eff}} \in \{500, 2000\}$ (160 runs total). Achieved IPS budgets track the targets (477 to 507 and 1915 to 2000 for the LSTM grid). Table 13 reports the results: contrasts are negative in all four architecture-target cells and statistically significant in three, and in no cell do gaps increase with dependence at matched information, matching the TCN result.

C.4 Beijing PM2.5 matched at a scarce budget

The main-text Beijing panel (Figure 2) matches the three preprocessing regimes at the budget implied by the fixed panel ($N_{\text{eff}} \approx 270$). Here the same panel is matched at the scarce target $N_{\text{eff}} = 100$, with the pilot

Table 13: Architecture generality at matched budgets: mean |Gap| per ρ and paired contrasts.

Architecture	Target	|Gap| by ρ				Contrast (95% CI)
		0.2	0.4	0.6	0.8	
LSTM	500	0.056	0.058	0.045	0.021	-0.034 (includes 0)
LSTM	2000	0.046	0.034	0.026	0.010	-0.036 $[-0.070, -0.002]$
Transformer	500	0.085	0.093	0.040	0.013	-0.072 $[-0.119, -0.024]$
Transformer	2000	0.046	0.024	0.018	0.009	-0.038 $[-0.056, -0.019]$

estimated on a 5,000-point segment at the deployment sequence scale. Pilot estimates $\widehat{\tau} = 37.0/41.8/46.8$ (raw/smooth6/smooth12) give allocated raw lengths 3704/4178/4676, all within 15% of the fixed $N = 4096$; achieved budgets are 101.8/101.3/99.5, within 2% of target. The accompanying fixed-N rerun reproduces the fixed-N panel (achieved $\widehat{N}_{\text{eff}}$ 114.0/99.6/87.9; |Gap| 0.096/0.029/0.031). Table 14 summarizes. Raw signed gaps on this panel are negative (test loss below train loss, reflecting nonstationarity across the chronological split), so magnitudes are reported, and the panel remains descriptive because smoothing changes the task. Missing PM2.5 values ($\approx 2.6\%$ of the series) are linearly interpolated before splitting; the fill is identical across the three regimes and both budget levels, so it cannot drive the between-regime or between-budget comparisons.

Table 14: Beijing PM2.5: fixed-N versus matching at the scarce target $N_{\text{eff}} = 100$ (deployment-scale pilot).

	Fixed $N = 4096$		Matched at $N_{\text{eff}} = 100$		
Regime	Achieved $\widehat{N}_{\text{eff}}$	|Gap|	Allocated N	Achieved $\widehat{N}_{\text{eff}}$	|Gap|
raw	114.0	0.096	3704	101.8	0.112
smooth6	99.6	0.029	4178	101.3	0.042
smooth12	87.9	0.031	4676	99.5	0.050

Comparing the two matched targets exposes a budget-sensitivity axis that fixed-N evaluation cannot show: the raw regime's gap falls from 0.112 at budget ≈ 100 to 0.016 at budget ≈ 270, while the smoothed arms remain essentially flat across the two budgets. Which preprocessing appears best therefore depends on the common budget level, and a fixed-N experiment pins every regime to one arbitrary slice of this axis.

C.5 Long-memory pilot diagnostic

The pilot-stage check re-estimates $\widehat{\tau}_{\text{int}}$ under an increasing maximum lag $M \in \{100, \ldots, 5000\}$ on the pilot sample. Long-memory series are ARFIMA$(0, d, 0)$ with fractional-differencing parameter d. For AR(1) with $\rho = 0.6$, the estimate is flat at 3.94 across the entire range (standard error 0.02): a plateau, so the protocol applies. For long-memory ARFIMA processes the estimate grows without plateauing: for $d = 0.30$ from 20.5 to 104.4 (a $5\times$ increase; standard error growing from 0.13 to 22.2), and for $d = 0.45$ from 75.3 to 1139.4 (a $15\times$ increase; standard error from 1.3 to 136.5). Figure 15 shows the curves. The diagnostic flags strong violations; weak long memory ($d = 0.10$ near-plateaus at 2.65) may require longer pilot horizons or dedicated estimators such as local-Whittle.

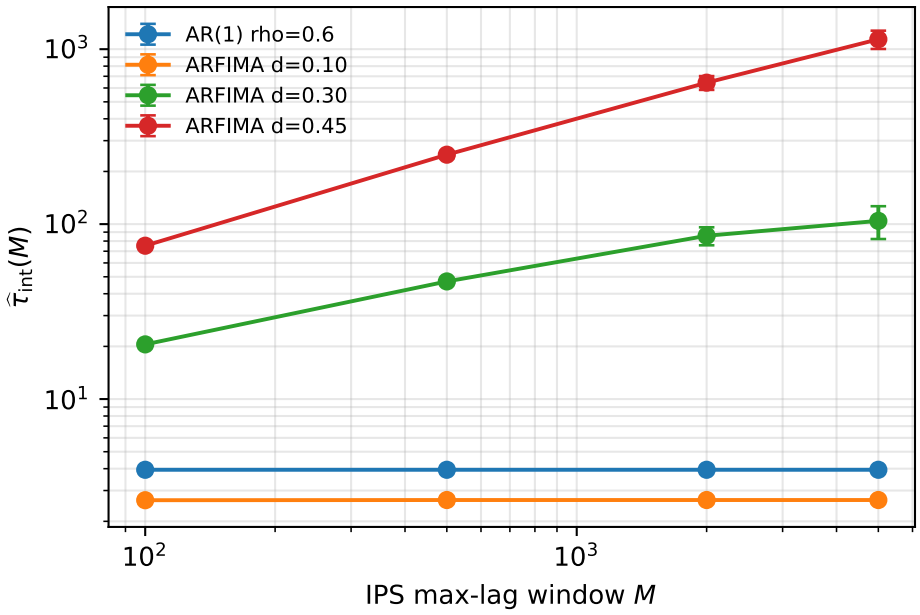


Figure 15: Pilot-stage mixing diagnostic: $\widehat{\tau}_{\text{int}}$ as a function of the maximum lag M. AR(1) plateaus; long-memory ARFIMA grows without plateau, flagging regimes where the matching protocol should not be applied.

D OMITTED PROOFS

This section provides full proofs aligned with the main-text pipeline: *(i) a blocking/coupling lemma for anchors under β-mixing; (ii) a generic dependent-to-i.i.d. reduction theorem (Theorem 4.2); (iii) scope lemmas for anchor/full empirical risk and block averaging; (iv) an i.i.d. Rademacher complexity bound for norm-controlled TCNs (Lemma 4.3); (v) the main architecture-aware baseline (Theorem 4.4) as a direct combination.*

D.1 Setup and notation

Let $(Z_t)_{t\geq 1}$ be a strictly stationary process on $\mathcal{Z}$. For $k \geq 0$, let $\beta(k)$ denote the (absolute regularity) β-mixing coefficient. We work under Assumption 3.1, restated here for convenience:

Assumption D.1 (Exponential β-mixing (Restatement of Assumption 3.1))**.** There exist constants $C_0, c_0 > 0$ such that for all $k \geq 0$,

$$\beta(k) \;\leq\; C_0 e^{-c_0 k}.$$

Fix a *delay* $d \geq 0$ and define the number of *anchors*

$$B \;=\; \left\lfloor \frac{m}{d+1} \right\rfloor, \qquad t_j \;=\; 1 + (j-1)(d+1), \quad j = 1, \ldots, B.$$

The *anchor sample* is $(Z_{t_1}, \ldots, Z_{t_B})$. For a loss $\ell : \mathcal{F} \times \mathcal{Z} \to [0,1]$ and predictor $f \in \mathcal{F}$, define

$$\mathcal{L}(f) \;=\; \mathbb{E}[\ell(f, Z_1)], \qquad \widehat{\mathcal{L}}_B^{\text{anc}}(f) \;=\; \frac{1}{B} \sum_{j=1}^{B} \ell(f, Z_{t_j}).$$

Let $\mathfrak{R}_B(\ell \circ \mathcal{F})$ denote the i.i.d. Rademacher complexity of the class $\{z \mapsto \ell(f, z) : f \in \mathcal{F}\}$ evaluated on B i.i.d. samples from the marginal distribution of Z_1.

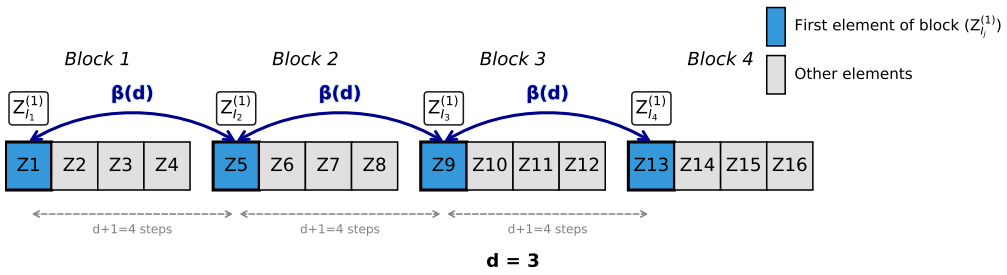


Figure 16: **Blocking with anchors.** We partition the dependent sequence into blocks of length $d+1$ and select one *anchor* per block (blue). Anchors are separated by $d+1$ time steps, so dependence decays with $\beta(d+1)$. Choosing $d \sim \log m$ yields $B \sim m/\log m$ anchors, so the final rate gains a $\sqrt{\log m}$ factor.

D.2 Proof of Lemma 4.1 (Blocking/Coupling Lemma)

Lemma D.2 (Restatement of Lemma 4.1)**.** *Let $(Z_t)_{t\geq 1}$ be strictly stationary and β-mixing. For anchors $(Z_{t_1}, \ldots, Z_{t_B})$ defined above,*

$$\left\| P_{Z_{t_1},\ldots,Z_{t_B}} - P_{Z_1}^{\otimes B} \right\|_{\mathrm{TV}} \;\leq\; (B-1)\,\beta(d+1) \;\leq\; B\,\beta(d+1).$$

Proof. Write $A_j = Z_{t_j}$ with $t_j = 1 + (j-1)(d+1)$. For $j = 1, \ldots, B$, define the intermediate measures

$$\mu_j \;:=\; P_{A_1,\ldots,A_j} \;\otimes\; \bigotimes_{k=j+1}^{B} P_{A_k}.$$

Then $\mu_B = P_{A_1,\ldots,A_B}$ and $\mu_1 = \bigotimes_{k=1}^{B} P_{A_k}$. By the triangle inequality,

$$\left\| P_{A_1,\ldots,A_B} - \textstyle\bigotimes_{k=1}^{B} P_{A_k} \right\|_{\mathrm{TV}} = \|\mu_B - \mu_1\|_{\mathrm{TV}} \leq \textstyle\sum_{j=1}^{B-1} \|\mu_{j+1} - \mu_j\|_{\mathrm{TV}}.$$

Moreover, tensoring both measures with the same product measure does not change total variation, hence

$$\|\mu_{j+1} - \mu_j\|_{\mathrm{TV}} = \left\| P_{A_1,\ldots,A_{j+1}} - P_{A_1,\ldots,A_j} \otimes P_{A_{j+1}} \right\|_{\mathrm{TV}}.$$

Now $\sigma(A_1, \ldots, A_j) \subseteq \mathcal{F}_{\leq t_j}$ and $\sigma(A_{j+1}) \subseteq \mathcal{F}_{\geq t_{j+1}} = \mathcal{F}_{\geq t_j + (d+1)}$, so these σ-algebras are separated by $d+1$. By the definition of β-mixing (absolute regularity),

$$\left\| P_{A_1,\ldots,A_{j+1}} - P_{A_1,\ldots,A_j} \otimes P_{A_{j+1}} \right\|_{\mathrm{TV}} \leq \beta(d+1).$$

Summing over $j = 1, \ldots, B-1$ yields

$$\left\| P_{A_1,\ldots,A_B} - \textstyle\bigotimes_{k=1}^{B} P_{A_k} \right\|_{\mathrm{TV}} \leq (B-1)\beta(d+1).$$

Finally, by stationarity $P_{A_k} = P_{Z_1}$ for all k, so $\bigotimes_{k=1}^{B} P_{A_k} = P_{Z_1}^{\otimes B}$. □

D.3 Proof of Theorem 4.2(Generic dependent-to-i.i.d. reduction)

Theorem D.3 (Generic anchor bound under β-mixing)**.** *Let $(Z_t)_{t\geq 1}$ be strictly stationary and β-mixing, and let $\ell : \mathcal{F} \times \mathcal{Z} \to [0,1]$. Fix $d \geq 0$ and define $B = \lfloor m/(d+1) \rfloor$ anchors as above. Then for any $\delta \in (0,1)$, with probability at least $1 - \delta - (B-1)\beta(d+1)$,*

$$\sup_{f\in\mathcal{F}} \left| \mathcal{L}(f) - \widehat{\mathcal{L}}_B^{\mathrm{anc}}(f) \right| \;\leq\; 2\,\mathfrak{R}_B(\ell \circ \mathcal{F}) \;+\; 3\sqrt{\frac{\log(2/\delta)}{2B}}.$$

Proof. Let $\widetilde{A} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_B)$ be an i.i.d. sample with marginals P_{Z_1}, and define

$$\widehat{\mathcal{L}}_B^{\mathrm{iid}}(f) \;=\; \frac{1}{B}\sum_{j=1}^{B} \ell(f, \widetilde{Z}_j).$$

Define the event $\mathcal{E} \subseteq \mathcal{Z}^B$ by

$$\mathcal{E} := \left\{ z_{1:B} \in \mathcal{Z}^B : \sup_{f\in\mathcal{F}} \left| \mathcal{L}(f) - \frac{1}{B}\sum_{j=1}^{B} \ell(f, z_j) \right| \le 2\,\mathfrak{R}_B(\ell \circ \mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2B}} \right\}.$$

By standard i.i.d. uniform convergence (symmetrization + concentration) for bounded losses in $[0,1]$,

$$\mathbb{P}_{\widetilde{A}}(\mathcal{E}) \;\ge\; 1 - \delta.$$

By Lemma 4.1,

$$\|P_A - P_{\widetilde{A}}\|_{\mathrm{TV}} \;\le\; (B-1)\beta(d+1),$$

and therefore for any event $\mathcal{E}$ we have $\mathbb{P}_A(\mathcal{E}) \ge \mathbb{P}_{\widetilde{A}}(\mathcal{E}) - \|P_A - P_{\widetilde{A}}\|_{\mathrm{TV}}$. Applying this to the above $\mathcal{E}$ gives

$$\mathbb{P}_A(\mathcal{E}) \;\ge\; 1 - \delta - (B-1)\beta(d+1).$$

On $\mathcal{E}$ (evaluated at $A = (Z_{t_1}, \ldots, Z_{t_B})$), we obtain exactly the claimed inequality. □

Before turning to the TCN complexity ingredient, we collect scope lemmas explaining why the theorem controls anchor risk rather than the ordinary full empirical risk. These lemmas are not needed to combine the main bound, but they clarify its interpretation and motivate the empirical anchor/block diagnostics.

Lemma D.4 (Anchor and full empirical risk: deterministic relation)**.** *Let*

$$\widehat{\mathcal{L}}_m(f) = \frac{1}{m}\sum_{t=1}^{m} \ell(f, Z_t)$$

be the full empirical risk computed on all m observations, and let

$$\widehat{\mathcal{L}}_B^{\mathrm{anc}}(f) = \frac{1}{B}\sum_{j=1}^{B} \ell(f, A_j)$$

be the anchor empirical risk computed on the B anchor points. Under stationarity and with loss $\ell \in [0,1]$:

1. *Both estimators are unbiased: $\mathbb{E}\left[\widehat{\mathcal{L}}_m(f)\right] = \mathbb{E}\left[\widehat{\mathcal{L}}_B^{\mathrm{anc}}(f)\right] = \mathcal{L}(f)$ for all $f \in \mathcal{F}$.*
2. *The empirical quantities satisfy the deterministic bound:*

$$\left|\widehat{\mathcal{L}}_m(f) - \widehat{\mathcal{L}}_B^{\mathrm{anc}}(f)\right| \le 1 - \frac{B}{m}.$$

Proof. By stationarity, $\mathbb{E}[\ell(f, Z_t)] = \mathcal{L}(f)$ for all t. Therefore,

$$\mathbb{E}\left[\widehat{\mathcal{L}}_m(f)\right] = \frac{1}{m}\sum_{t=1}^{m} \mathbb{E}[\ell(f, Z_t)] = \mathcal{L}(f),$$

and similarly for $\widehat{\mathcal{L}}_B^{\mathrm{anc}}(f)$.

Let $\mathcal{T} = \{t_1, \ldots, t_B\} \subset \{1, \ldots, m\}$ denote the anchor indices. Decompose

$$\widehat{\mathcal{L}}_m(f) = \frac{1}{m}\sum_{t=1}^{m} \ell(f, Z_t) = \frac{1}{m}\left(\sum_{j=1}^{B} \ell(f, A_j) + \sum_{t\notin\mathcal{T}} \ell(f, Z_t)\right).$$

Let $\widehat{\mathcal{L}}^{\mathrm{non}}_{m-B}(f) = \frac{1}{m-B}\sum_{t\notin\mathcal{T}} \ell(f, Z_t)$ denote the average over non-anchor points (when $m > B$). Then:

$$\widehat{\mathcal{L}}_m(f) = \frac{B}{m}\cdot\widehat{\mathcal{L}}^{\mathrm{anc}}_B(f) + \frac{m-B}{m}\cdot\widehat{\mathcal{L}}^{\mathrm{non}}_{m-B}(f).$$

Since both averages lie in $[0,1]$,

$$\begin{aligned}\left|\widehat{\mathcal{L}}_m(f) - \widehat{\mathcal{L}}^{\mathrm{anc}}_B(f)\right| &= \left|\frac{B}{m}\cdot\widehat{\mathcal{L}}^{\mathrm{anc}}_B(f) + \frac{m-B}{m}\cdot\widehat{\mathcal{L}}^{\mathrm{non}}_{m-B}(f) - \widehat{\mathcal{L}}^{\mathrm{anc}}_B(f)\right| \\ &= \frac{m-B}{m}\left|\widehat{\mathcal{L}}^{\mathrm{non}}_{m-B}(f) - \widehat{\mathcal{L}}^{\mathrm{anc}}_B(f)\right| \\ &\le \frac{m-B}{m}\cdot 1 = 1 - \frac{B}{m}.\end{aligned}$$

□

Corollary D.5 (Anchor-vs-full empirical risk: no small-gap guarantee without extra structure)**.** *With $B = \lfloor m/(d+1)\rfloor$, Lemma D.4(ii) implies*

$$\left|\widehat{\mathcal{L}}_m(f) - \widehat{\mathcal{L}}^{\mathrm{anc}}_B(f)\right| \le 1 - \frac{B}{m}.$$

For the logarithmic delay choice $d = \Theta(\log m)$, we have $\frac{B}{m} = \Theta(1/\log m)$ and thus $1 - \frac{B}{m} = 1 - \Theta(1/\log m)$, which is not *small in general. Therefore, our theory is stated directly for the anchor empirical risk; relating it to the full empirical risk would require additional assumptions (e.g., stability/smoothness of losses within blocks) or a different estimator (e.g., block-averaged losses).*

D.4 A block-averaged empirical risk

The one-point anchor estimator is the simplest empirical object compatible with the blocking proof, but it is not the only one. A retained block can average several consecutive losses before the next gap. This uses more of the observed loss sequence while preserving the same block-level coupling logic.

Lemma D.6 (Block-averaged empirical risk under β-mixing)**.** *Let $(Z_t)_{t\ge1}$ be strictly stationary and β-mixing, with losses $\ell(f, Z_t) \in [0,1]$. Fix a retained block length $a \ge 1$ and a gap length $g \ge 0$. Define*

$$s_j = 1 + (j-1)(a+g), \qquad I_j = \{s_j, \dots, s_j + a - 1\}, \qquad B = \left\lfloor \frac{m}{a+g} \right\rfloor,$$

and the block-averaged empirical risk

$$\widehat{\mathcal{L}}^{\mathrm{blk}}_{a,g}(f) = \frac{1}{Ba}\sum_{j=1}^{B}\sum_{t\in I_j}\ell(f, Z_t).$$

Let $V_j = (Z_t)_{t\in I_j}$ and define the block-loss class

$$\Phi_a = \left\{ v = (z_1, \dots, z_a) \mapsto \frac{1}{a}\sum_{r=1}^{a}\ell(f, z_r) :\ f \in \mathcal{F} \right\}.$$

Then $\mathbb{E}[\widehat{\mathcal{L}}^{\mathrm{blk}}_{a,g}(f)] = \mathcal{L}(f)$ for each fixed f. Moreover, for any $\delta \in (0,1)$, with probability at least $1-\delta-(B-1)\beta(g+1)$,

$$\sup_{f\in\mathcal{F}}\left|\mathcal{L}(f) - \widehat{\mathcal{L}}^{\mathrm{blk}}_{a,g}(f)\right| \le 2\,\mathfrak{R}_B(\Phi_a) + 3\sqrt{\frac{\log(2/\delta)}{2B}}.$$

Finally, for every f,

$$\left|\widehat{\mathcal{L}}_m(f) - \widehat{\mathcal{L}}^{\mathrm{blk}}_{a,g}(f)\right| \le 1 - \frac{Ba}{m}.$$

Proof. Stationarity gives $\mathbb{E}[\ell(f, Z_t)] = \mathcal{L}(f)$ for every t, hence the block average is unbiased. For coupling, the last index of V_j is $s_j + a - 1$, while the first index of V_{j+1} is $s_j + a + g$. Thus the two block σ-algebras are separated by $g+1$ steps in the windowed example process. The same telescoping argument as Lemma 4.1 gives

$$\left\| P_{V_1,\dots,V_B} - \bigotimes_{j=1}^{B} P_{V_j} \right\|_{\mathrm{TV}} \le (B-1)\beta(g+1).$$

On the coupled i.i.d. block sample, apply the standard bounded-loss Rademacher bound to the class Φ_a. Transferring the event back through total variation yields the displayed probability and bound.

For the deterministic relation, let $T = \cup_{j=1}^{B} I_j$ be the retained indices and let $\widehat{\mathcal{L}}^{\mathrm{disc}}_{m-Ba}(f)$ be the average over discarded indices when $m > Ba$. Then

$$\widehat{\mathcal{L}}_m(f) = \frac{Ba}{m}\widehat{\mathcal{L}}^{\mathrm{blk}}_{a,g}(f) + \frac{m-Ba}{m}\widehat{\mathcal{L}}^{\mathrm{disc}}_{m-Ba}(f).$$

Since both averages lie in $[0,1]$,

$$\left|\widehat{\mathcal{L}}_m(f) - \widehat{\mathcal{L}}^{\mathrm{blk}}_{a,g}(f)\right| = \frac{m-Ba}{m}\left|\widehat{\mathcal{L}}^{\mathrm{disc}}_{m-Ba}(f) - \widehat{\mathcal{L}}^{\mathrm{blk}}_{a,g}(f)\right| \le 1 - \frac{Ba}{m}.$$

□

Remark D.7 (What block averaging does not claim). Lemma D.6 treats each retained block as one approximately independent unit. It therefore does not give a $1/\sqrt{Ba}$ rate and does not make the ordinary full empirical risk independent. Its role is to provide a computable block-compatible empirical risk that can be much closer to the full empirical risk than one-point anchors, at the cost of a concentration term governed by the number of retained blocks B.

Lemma D.8 (Anchor–full risk under within-block smoothness)**.** *Let $s = d+1$, $B = \lfloor m/s \rfloor$, and $t_j = 1 + (j-1)s$ be the anchor in block $I_j = \{t_j, \dots, t_j + s - 1\}$ for $j = 1, \dots, B$. Assume $\ell(f, Z_t) \in [0,1]$ and that for a fixed f there is a semimetric $d_{\mathcal{Z}}$ and constant $L_{\mathrm{blk}}(f)$ such that for every complete block and every $t \in I_j$,*

$$\left|\ell(f, Z_t) - \ell(f, Z_{t_j})\right| \le L_{\mathrm{blk}}(f)\, d_{\mathcal{Z}}(Z_t, Z_{t_j}).$$

Define the within-block dispersion

$$\overline{\Delta}_d(f) = \frac{1}{Bs}\sum_{j=1}^{B}\sum_{t \in I_j} d_{\mathcal{Z}}(Z_t, Z_{t_j}).$$

Then, with $r = m - Bs < s$,

$$\left|\widehat{\mathcal{L}}_m(f) - \widehat{\mathcal{L}}^{\mathrm{anc}}_B(f)\right| \le L_{\mathrm{blk}}(f)\,\overline{\Delta}_d(f) + \frac{2r}{m} \le L_{\mathrm{blk}}(f)\,\overline{\Delta}_d(f) + \frac{2d}{m}.$$

The same statement holds uniformly over a class $\mathcal{F}$ if the displayed within-block condition holds for all $f \in \mathcal{F}$ with a common constant L_{blk}.

Proof. Write the full empirical risk as the contribution from the B complete blocks plus the remainder:

$$\widehat{\mathcal{L}}_m(f) = \frac{Bs}{m}\cdot\frac{1}{Bs}\sum_{j=1}^{B}\sum_{t \in I_j}\ell(f, Z_t) + \frac{1}{m}\sum_{t=Bs+1}^{m}\ell(f, Z_t).$$

Because losses are bounded by 1, the remainder contributes at most r/m to the absolute difference, and the coefficient mismatch between the complete-block average and its weight Bs/m contributes another at most r/m. For the complete blocks, subtracting the anchor average gives

$$\left|\frac{1}{Bs}\sum_{j=1}^{B}\sum_{t \in I_j}\ell(f, Z_t) - \frac{1}{B}\sum_{j=1}^{B}\ell(f, Z_{t_j})\right| \le \frac{1}{Bs}\sum_{j=1}^{B}\sum_{t \in I_j}\left|\ell(f, Z_t) - \ell(f, Z_{t_j})\right| \le L_{\mathrm{blk}}(f)\overline{\Delta}_d(f).$$

Combining the two bounds and using $r < s = d+1$ yields the claim. □

D.5 Proof of Lemma 4.3 (i.i.d. Rademacher bound for norm-controlled TCNs)

Lemma D.9 (Norm-controlled TCN Rademacher bound)**.** *Assume inputs are bounded: $\|x\|_F \le B_x$ almost surely. Let $\mathcal{F}_{D,p,R}$ be the class of depth-D causal TCNs with kernel size p, ReLU activations, and layer-wise $\ell_{2,1}$ norm bounds $\|W^{(\ell)}\|_{2,1} \le M^{(\ell)}$ with product budget $R = \prod_{\ell=1}^{D} M^{(\ell)}$. Then, up to the architecture/logarithmic factor*

$\Lambda_{\mathrm{TCN}}(B, \mathbf{C}_{\mathrm{ch}}, n, q, p)$ *made explicit by Ledent et al. (2021), there is no linear dependence on the raw sequence length and for i.i.d. samples of size* B*,*

$$\mathfrak{R}_B^{\mathrm{vec}}(\mathcal{F}_{D,p,R}) \;\le\; \Lambda_{\mathrm{TCN}}(B, \mathbf{C}_{\mathrm{ch}}, n, q, p)\, \frac{R\, B_x \sqrt{D \log(2p)}}{\sqrt{B}}.$$

Proof. We do not rederive the full multilayer convolutional covering argument. Instead, we verify that our 1D causal convolution class satisfies the norm-control hypotheses used in the convolutional Rademacher analysis of Ledent et al. (2021), and then invoke their weight-sharing bound as the i.i.d. architectural ingredient.

Step 1: Local patch control under $\ell_{2,1}$ constraints. For a convolutional layer $\phi_W(x) = W * x$ with weight tensor $W \in \mathbb{R}^{C_{\mathrm{out}} \times C_{\mathrm{in}} \times p}$, the filter-group norm is

$$\|W\|_{2,1} = \sum_{j=1}^{C_{\mathrm{out}}} \|W_{j,:,:}\|_F,$$

where $W_{j,:,:} \in \mathbb{R}^{C_{\mathrm{in}} \times p}$ is the j-th output filter.

Local claim. For each output channel j and spatial position t, the convolution computes $(W * x)_{j,t} = \langle W_{j,:,:}, x_{:,t-p+1:t} \rangle$. By Cauchy–Schwarz,

$$|(W * x)_{j,t}| \le \|W_{j,:,:}\|_F \cdot \|x_{:,t-p+1:t}\|_F.$$

This is the only norm-control verification needed to match our notation to the convolutional covering result. We do *not* claim that overlapping convolutions satisfy a global bound $\|W * x\|_F \le \|W\|_{2,1} \|x\|_F$ without additional factors: summing over overlapping patches can introduce a $\sqrt{p}$ factor. The convolutional covering theorem handles repeated local filter applications through weight sharing rather than dense Toeplitz expansion.

Step 2: Invoke the convolutional weight-sharing bound. For convolutional layers, each filter is applied at many positions. A dense Toeplitz view would count repeated weights many times and introduce spurious dependence on the number of filter applications. We use the multilayer convolutional covering theorem of Ledent et al. (2021, Theorem 5, specialized through Proposition 6). In their notation, our causal 1D convolution is a one-dimensional convolutional layer; the row/filter matrix is the collection of tensors $W_{j,:,:}$ above; the patch diameter is controlled by B_x; sample–patch–channel triples enter logarithmically in B, q, and $\mathbf{C}_{\mathrm{ch}}$; and the output dimension is the fixed forecasting dimension n. Their one-layer $\ell_{2,1}$ cover counts each filter once, and their multilayer theorem combines those covers with the norm-product/depth argument of Golowich et al. (2018). Applying that result to our causal 1D architecture and local filter convention gives

$$\mathfrak{R}_B^{\mathrm{vec}}(\mathcal{F}_{D,p,R}) \le \Lambda_{\mathrm{TCN}}(B, \mathbf{C}_{\mathrm{ch}}, n, q, p)\, \frac{RB_x \sqrt{D \log(2p)}}{\sqrt{B}},$$

where Λ_{TCN} suppresses logarithmic factors in B and fixed architecture constants such as channel widths, context length, and output dimension. No factor proportional to raw sequence length appears; context-window dependence enters through B_x and fixed patch-cover factors. □

Remark D.10 (Constants and logarithmic factors). We no longer claim an explicit universal value such as $4\sqrt{2}$ for this lemma. The constants and logarithmic terms come from the cited i.i.d. convolutional covering theorem; the main text keeps only the architectural scaling needed for the dependent-data baseline.

Remark D.11 (Why $\ell_{2,1}$ norm, not spectral norm). For general convolution operators, the spectral norm is *not* bounded by the $\ell_{2,1}$ filter-group norm: a uniform kernel $W = [1, \ldots, 1] \in \mathbb{R}^p$ has $\|W\|_{2,1} = \sqrt{p}$ but operator norm p at DC frequency. We therefore do not claim spectral control or a global no-overlap Lipschitz bound; $\ell_{2,1}$ is only the local filter norm used in the convolutional covering argument of Ledent et al. (2021).

D.6 Technical lemma: Lipschitz loss (squared loss under bounded outputs)

Theorem 4.4 uses vector-valued Lipschitz contraction (Maurer, 2016): for an L-Lipschitz loss in Euclidean norm, $\mathfrak{R}_B(\ell \circ \mathcal{F}) \le C_{\mathrm{vc}} L\, \mathfrak{R}_B^{\mathrm{vec}}(\mathcal{F})$ up to a universal constant C_{vc}. For squared loss, this requires bounded predictions (or clipping).

Lemma D.12 (Squared loss is Lipschitz on an ℓ_2-ball)**.** *Assume $\|y\|_2 \le B_y$ and $\|\hat{y}\|_2, \|\hat{y}'\|_2 \le B_f$. Then $\ell(\hat{y}, y) = \|\hat{y} - y\|_2^2$ satisfies*

$$|\ell(\hat{y}, y) - \ell(\hat{y}', y)| = |\langle \hat{y} - \hat{y}',\, \hat{y} + \hat{y}' - 2y \rangle| \le 2(B_f + B_y)\, \|\hat{y} - \hat{y}'\|_2.$$

Proof. Let $\hat{y}, \hat{y}', y \in \mathbb{R}^k$. Using the polarization identity,

$$\|\hat{y} - y\|_2^2 - \|\hat{y}' - y\|_2^2 = \langle \hat{y} - \hat{y}',\, \hat{y} + \hat{y}' - 2y \rangle.$$

Hence, by Cauchy–Schwarz and the assumed bounds,

$$\begin{aligned} \big|\|\hat{y} - y\|_2^2 - \|\hat{y}' - y\|_2^2\big| &\le \|\hat{y} - \hat{y}'\|_2 \cdot \|\hat{y} + \hat{y}' - 2y\|_2 \\ &\le \|\hat{y} - \hat{y}'\|_2 \cdot \big(\|\hat{y}\|_2 + \|\hat{y}'\|_2 + 2\|y\|_2\big) \\ &\le 2(B_f + B_y)\, \|\hat{y} - \hat{y}'\|_2. \end{aligned}$$

□

D.7 Proof of Theorem 4.4 (Architecture-aware baseline under exponential β-mixing)

Theorem D.13 (Restatement of Theorem 4.4)**.** *Assume exponential β-mixing (Assumption 3.1), bounded inputs/targets/outputs (Assumption 3.3), and Lipschitz loss (Assumption 3.5). For squared loss, use the normalized loss convention from Section 4 and rescale the final inequality back to MSE units. Let $\mathcal{F}_{D,p,R}$ be the norm-controlled TCN class in Lemma 4.3, and let $B = \lfloor m/(d+1) \rfloor$ be the number of anchors. Then with probability at least $1 - \delta - (B-1)\beta(d+1)$,*

$$\sup_{f \in \mathcal{F}_{D,p,R}} \Big| \mathcal{L}(f) - \widehat{\mathcal{L}}_B^{\mathrm{anc}}(f) \Big| \;\le\; C' \, \frac{L\, \Lambda_{\mathrm{TCN}}\, R\, B_x \sqrt{D\, \log(2p)}}{\sqrt{B}} \;+\; 3\sqrt{\frac{\log(2/\delta)}{2B}}$$

where $\Lambda_{\mathrm{TCN}} = \Lambda_{\mathrm{TCN}}(B, \mathbf{C}_{\mathrm{ch}}, n, q, p)$ is the architecture/log factor from Lemma 4.3, and C' is universal.

Proof. Apply Theorem 4.2 with $\mathcal{F} = \mathcal{F}_{D,p,R}$ to obtain that with probability at least $1 - \delta - (B-1)\beta(d+1)$,

$$\sup_{f \in \mathcal{F}_{D,p,R}} \big| \mathcal{L}(f) - \widehat{\mathcal{L}}_B^{\mathrm{anc}}(f) \big| \le 2\, \mathfrak{R}_B(\ell \circ \mathcal{F}_{D,p,R}) + 3\sqrt{\frac{\log(2/\delta)}{2B}}.$$

By vector-valued Lipschitz contraction (Maurer, 2016) (valid for L-Lipschitz losses up to a universal constant; for squared loss use Lemma D.12 after bounding/clipping outputs),

$$\mathfrak{R}_B(\ell \circ \mathcal{F}_{D,p,R}) \;\le\; C_{\mathrm{vc}} L\, \mathfrak{R}_B^{\mathrm{vec}}(\mathcal{F}_{D,p,R}).$$

Then apply Lemma 4.3 to obtain

$$\mathfrak{R}_B(\ell \circ \mathcal{F}_{D,p,R}) \le C_{\mathrm{vc}}\, L \cdot \Lambda_{\mathrm{TCN}} \frac{R\, B_x \sqrt{D\, \log(2p)}}{\sqrt{B}}.$$

Combining and absorbing C_{vc} and the remaining universal constants into C' yields the stated inequality.

Rate under exponential mixing. If $\beta(k) \le C_0 e^{-c_0 k}$ and we choose $d = \lceil (\log m)/c_0 \rceil$, then $B = \Theta(m/\log m)$ and

$$B\, \beta(d+1) \;\le\; \Theta\Big(\frac{m}{\log m}\Big) \cdot C_0 e^{-c_0(d+1)} \;=\; O\Big(\frac{1}{\log m}\Big),$$

while the leading complexity term scales like

$$\frac{1}{\sqrt{B}} = \Theta\Big(\sqrt{\frac{\log m}{m}}\Big),$$

giving the rate stated in the main text. □

Remark D.14 (Faster-decaying residual via a larger delay). The coupling slack $O(1/\log m)$ decays slowly. Choosing $d = \lceil \kappa(\log m)/c_0 \rceil$ with $\kappa > 1$ yields $B\, \beta(d+1) = O\big(m^{1-\kappa}/\log m\big)$ while shrinking the anchor count B only by the constant factor κ, so the leading complexity term keeps the rate $\sqrt{(\log m)/m}$ up to constants. We state the theorem with $\kappa = 1$ for simplicity.